Module: **M6. Video analysis** **Final exam**
Date: May 2, 2019 **Time: 2h**
Teachers: Montse Pardàs, Ramon Morros, Xavier Giró, Javier Ruiz, Josep Ramon Casas.

- Books, lecture notes, calculators, phones, etc. are not allowed.
- All sheets of paper should have your name.
- All results should be demonstrated or justified.

BEGIN EACH OF THE 4 PARTS OF THE EXAM IN A NEW SHEET OF PAPER

Part 1

Question 1 (Shot segmentation): **2 Point**

In order to separate a video sequence in video shots, the FD function is computed, based of the difference between every two successive frames (Frame Difference) is computed. Figure 1 shows the result FD function obtained for frames 0 to 120 for a given video, which is composed of 4 video shots corresponding to frames 0-30, 31-67, 68-86 and 87-120. Another possibility is to compute the DFD (Displaced Frame Difference) after a Motion estimation algorithm is applied. The result DFD function in this case is shown in Figure 2.

a) How could you separate the different shots with a simple procedure? Which would be the result in both cases?
Thresholding the FD or DFD Energy Functions. In the first case we would obtain many False Alarms in the last shot if we use a Threshold between 2 and 3. In the second case the result would be correct for a threshold between 1,5 and 3.

b) Explain a possible reason for the differences between shots in the FD energy function.
The first and the forth shot contain more moving objects, or there is strong camera motion.

c) Explain a possible reason for the differences between FD and DFD energy functions.
Motion compensation reduces the DFD when frames belong to the same shot.

d) Shot detection can be achieved by segmentation of the functions in Figure 1 or 2. Propose and justify a robust algorithm to produce this segmentation.
Simplification of the 1D function, marker detection and region growing.

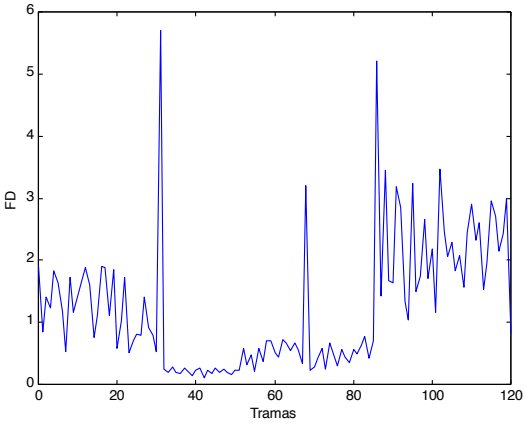


Figure 1: FD function

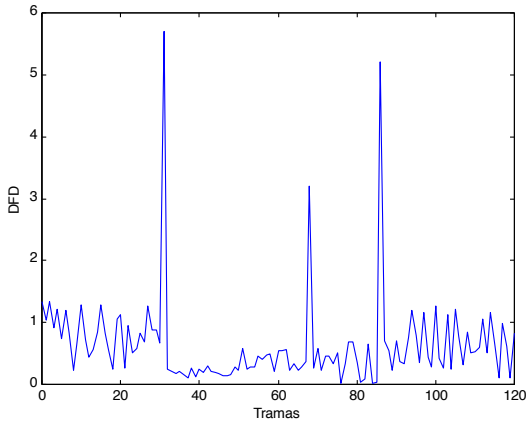


Figure 2: DFD function

Question 2 (Foreground segmentation): **2 Point**

We wish to insert real actors in a computer generated scenario. For this aim, a video with the actors is recorded in front of a blue screen, in such a way that foreground can be easily segmented and superimposed on the generated scene.
Due to illumination changes, the background does not appear as a unique color, and we decide to model the background color distribution with a probabilistic model.

a) A training image T(x,y) is provided, with pixels labeled as Foreground (F) or Background (B). Write the equations that you would use to estimate the background model using only this training image. How would you classify each pixel of the video sequence in Foreground or Background, for the whole video?
We can estimate a Gaussian model, with the same paràmetres for all píxels.
We need the mean of pixels labeled as B in T(x,y). If N if the number of these pixels: $m_B = \frac{\sum T(x,y)}{N}$
And the variance of these pixels: ${\sigma_B}^2 = \frac{\sum (T(x,y)-m_B)^2}{N}$
To classify a píxel, we assign F if $|I(x,y) - m_B| > \lambda\sigma_B$, and B otherwise.

b) No Training image is provided, and the blue screen suffers variations within an image due to illumination. Is it possible to compute an accurate background model for each pixel using the fist M frames? If possible, how would you do it? Under which circumstances do you think the model would be accurate enough?
We can compute the median of each pixel. The model will be accurate for those pixels which have observed more background than foreground in these M frames.

c) If the illumination produces strong shadows, propose a coordinate space different from RGB which might reduce the false foreground detections, together with a global strategy to follow.
We could use HSV. We first would detect F pixels and then for those pixels detected as F we would verify if they are shadows or not. A shadow pixel will have lower values of Value and Saturation than the corresponding background, and a similar Hue.

Question 3 (Contour tracking): **1 Point**

In Active Shape Models the object that we wish to find or track is described by a set of landmarks which are defined for a specific class of objects. To build the Shape Model these landmarks are firsts extracted from a set of training images.

a) Which are the main steps of the Shape Model Construction procedure?
See Slide 23

b) How can you determine if a new set of landmarks corresponds to the shape of an object of the class learned?
See slide 21. We first need to align the shape to the mean shape of the object class. Then we extract the shape parameters b of our new shape. To determine if the shape belongs to the object class we check if
$|b_j| < \beta\sqrt{\lambda_j}$

Part 2

Question 4 (Motion estimation): **2 Point**
Let's suppose a frame-level motion estimation technique using an affine model. The iterative solution is given by:

$$p_i^{k+1} = p_i^k - 2\mu \sum_{\vec{r}\in\mathbf{R}} DFD(\vec{r}, t, p_i^k)\nabla I(\vec{r} - \vec{D}(\vec{r}, p_i^k), t - \Delta t)\frac{\partial}{\partial p_i}\left(\vec{D}(\vec{r}, p_i^k)\right)$$

a) Describe the terms that appear in the equation.
b) Explain how the terms inside the sum can be computed.

a) p_i^{k+1}: motion parameters at iteration k+1
p_i^{k}: motion parameters at iteration k
DFD(r, t, p_i^k): Displaced frame difference
μ : step size in gradient descent (scalar)
∇I(…): spatial gradient of the motion compensated image, using the motion model parameters at iteration *k*.
δ(…)/δp_i: **derivative** of the optical flow with respect to the motion model parameters at iteration *k*

b) DFD: Difference between current frame and the motion compensated image, using the motion model parameters at iteration *k*
∇I(…): Apply sobel operator to the motion compensated image
δ(…)/δp_i: Compute the derivative of the affine model equations with respect to each one of the 6 model parameters

Question 5 (Motion estimation): **1 Point**
The Lucas-Kanade optical flow estimation is given by:
$V = (A^T A)^{-1} A^T \boldsymbol{b}$
Explain in which points of the image is it possible to obtain a reliable motion estimate.

a) Qualitatively, explaining how the regions around the point look like.
b) Quantitatively, stating the mathematical conditions that these points must fulfill.

a) A good estimate can be obtained in corners or textured regions
b) (A^TA) is invertible, its eigenvalues are not too small and are similar in value.

Question 6 (Tracking): **2 Point**

Tracking with particle filters. Explain the importance sampling method, why is it necessary and give the formula for E[X] in this case.

See slide 20 of MCV_M6_L6_-_ParticleFilters

$$E[X] \approx \frac{1}{N}\sum_{i=1}^{N} w^i f\left(x^i\right) \qquad w^i = \frac{p(x^i)}{q(x^i)}$$

Part 3

Question 7 (Model based tracking): **1 Point**

1) Explain the roles in generative approaches of the main elements of model-based tracking, i.e. Model, Estimator and Data (Observations)
2) How can each one of these three elements be improved in order to increase the accuracy of model based tracking?

1) Generative model-based tracking is an analysis-by-synthesis approach using a Human Body Model (HBM). The modelling phase involves the HBM definition, likelihood and matching functions. An estimator is used in the estimation phase in order to find the most likely pose according to the observations. The data (observations) is used to assess the hypothesis synthesized by the estimator and, therefore, some kind of metrics (or matching functions) should be defined for this assessment-
2) The model can be improved by increasing its precision or its flexibility. For instance, parts-based models part-based models allow for qualitative descriptions of visual appearance, and are suitable for generic recognition problems can be seen as a simpler and more flexible approach to track human bodies in a more generic way, as they are suitable for generic recognition problems (even to track non-human objects).
   The estimator can be improved if the model definition gets closer to the data. One case is Shotton's approach where fast and reliable estimation of the human body pose is obtained from images, by training a random forest classifier as estimator.
   Finally, richer data observations may improve the assessment of the hypothesis or the accuracy and precision of the estimation, as it was the case with 3D (RGB+D) data captured by commercial depth sensors

Question 8 (Gesture Recognition): **1 Point**

Enumerate four image features that are used to recognize **static** gestures from images in different gesture recognition systems.
2D color histograms, 2D sift, 2D/3D shape context, 3D Viewpoint Feature Histograms (VFH), 3D Oriented Radial Distribution (ORD), learned VGG, learned ResNet, etc.

Question 9 (Gesture Recognition): **1 Point**

Enumerate four video features that are used to recognize **dynamic** gestures from videos in different gesture recognition systems.
Motion Energy Image (MEI), Motion History Image (MHI), optical flow, trajectories, learned C3D, etc.

Part 4

Question 10 (Deep Learning Architectures for Video): **1 Point**

Consider a very simple neural network for grayscale image classification with a single layer convolutional layer that contains the following and single 2D filter:

0	3	0
3	6	3
0	3	0

A new task of grayscale video classification is defined for a video dataset whose frames are very similar to the images used to learned the 2D filter. For this reason, instead of training a new convolutional network from scratch, the parameters of the 2D filter want to be adapted for C3D filters of temporal depth 3.

What would be the shape and weights of the new C3D filter ?

Solution

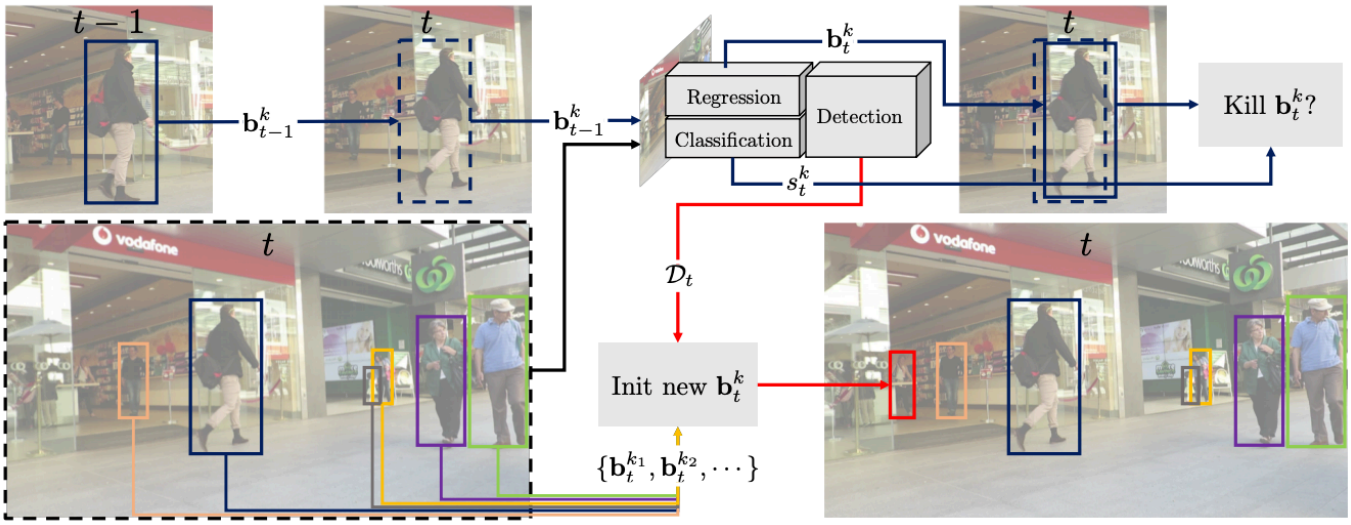
0	1	0
1	2	1
0	1	0

0	1	0
1	2	1
0	1	0

0	1	0
1	2	1
0	1	0

Question 11 (Deep Video Object Tracking): **1 Point**

Given the scheme of the *Tracktor* model (Bergmann et al, 2019):



a) What is the input of the *Regression* block?

The input are the RoI pooled features over the bounding box in the previous frame. The output are the coordinates of the regressed bounding boxes over the current frame.

b) What block is responsible of deciding that an object being tracked is no longer visible?

The Classification block, when the classification score is below a certain threshold.

c) How are new objects discovered?

The object detector provides a set of object detections D_t at frame t (not only the ones tracked from t-1). This allows initializing new tracks if none of the detections has an IoU larger than a certain threshold.

Question 12 (Deep Video Object Segmentation): **1 Point**

a) What does online learning in the context of video object segmentation as used, for example, by the OSVOS model?

The parameters of the model are adapted based on the mask of the object to track.

b) How can the temporal recurrence be considered in video object segmentation models?

By using recurrent neural networks with a hidden state maintained with learned weights (eg. RVOS), and by feeding as input the mask(s) predicted in the previous frame(s) (eg. LucidTracker).

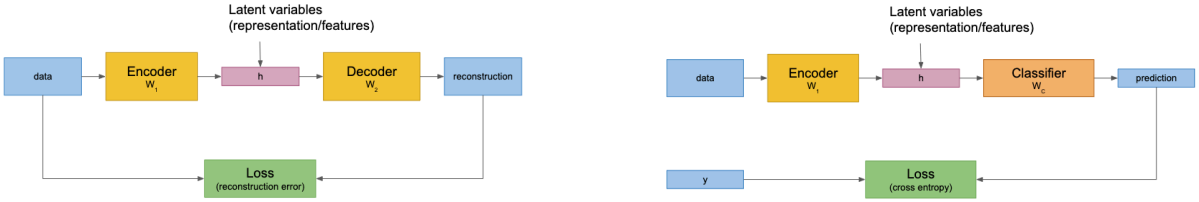
Question 13 (Self-supervised Learning from Video Sequences): **1 Point**

a) Define self-supervised learning.

Self-supervised learning is a form of unsupervised learning where the data provides the supervision. A **surrogate task** must be invented by withholding a part of the unlabeled data and training the NN to predict it.

b) Motivate the interest of training an autoencoder by in the scenario in which few labels are available for the final task. Base your argumentation with the necessary graphical schemes.

Solution:

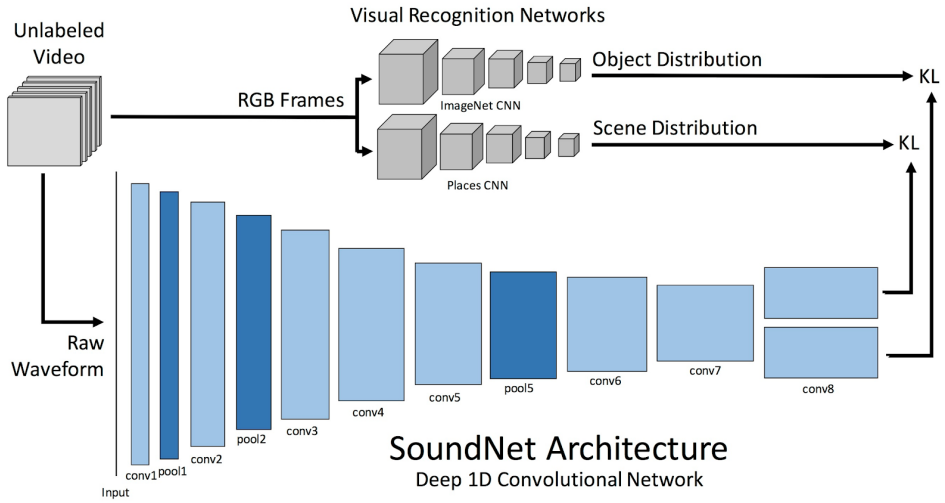


c) Describe a self-supervised task on video frames based on temporal verification.

- Temporal order of frames is exploited as the supervisory signal for learning.
- Train a network to detect which of the video sequences contains frames in the wrong order.
- Sort the sequence of frames.
- Predict whether the video moves forward or backward.

Question 14 (Self-supervised Audiovisual Learning): **1 Point**

Given the following scheme extracted from Soundnet (Aytar et al, 2016),



a) Which features are learnt?

The audio features.

b) Discuss the type of supervision exploited.

The visual recognition networks were trained in a fully supervised way. Nevertheless, the audio network is trained without any annotation of the audio, as there is an automatic transfer of (weak) labels from vision to audio.