Module: **M4. Video analysis** **Final exam**
Date: February 18th, 2016 **Time: 2h30**
Teachers: Montse Pardàs, Ramon Morros, David Varas, Constantine Butakoff, Josep Ramon Casas, Javier Ruiz, Jordi González, Xavier Giró.

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

Part 1 Montse Pardàs (2 points)

Question 1 **(1 point)**
Assuming you have just two frames of static background available, explain how you can model the background for foreground subtraction. Explain also a possible process for updating this model once the foreground subtraction takes place.

Answer: A gaussian model for each pixel can be used, assigning the mean color between the two frames of each pixel to the mean of its corresponding gaussian model. The variance can be estimated with an histogram of the frame difference between the two frames. With this histogram we can set the thresholds for foreground detection.
Running gaussian average can be used for the updating. See slide in 4.2, page 5.

Question 2 **(0.5 point)**
Describe different ways to define connectivity when performing spatio-temporal segmentation in video sequences

Answer: See slide in 4.2, page 28

Question 3 **(0.5 point)**
Explain different ways to include motion features when performing video sequence segmenation

Answer: Motion can be used as a homogeneity feature in different ways:
- Compute a dense optical flow and merge pixels which are homogeneous in motion
- A color or texture based segmentation can be computed, and regions with homogeneous motion can be merged

Part II Ramon Morros (1 point)

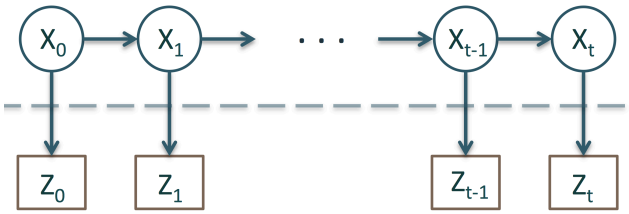
Question 1 **(1 point)**
The Lucas-Kanade method to compute optical flow assumes small motion.
a) When deriving the L-K equation to compute optical flow from the brightness constancy equation, explain in which step is this assumption introduced.
b) Explain some variation of the L-K method so that it can be used in the case of large motion.

Answer:
a) The L-H equation is obtained by taking a Taylor expansion of the brightness constancy equation and keeping first order terms only. Higher order terms will be null only if motion is small.
b) Iterative and multiresolution L-K. See slides (pp 63-66) for a complete explanation.

Part III **David Varas** **(2 points)**

Question 1 **(0.5 point)**

Consider the following scheme:



a) Define state and measurement and relate these concepts with the scheme.
State (x_i) is formed by the true characteristics that represent the tracked object.
Measurement (z_i) is what we can measure of the image at a given time.

b) Regarding this definition, explain the main conceptual difference between the areas above and below the dashed line in terms of data availability.
The area above the dashed line represents the part of the problem that are not observable, while the area under the dashed line represent the part of the problem to which we have access.

c) Explain how states and measurements are related with the tracking steps (prediction and update).
Prediction is performed before the measurement arrives. Only the previous state is used to compute the current state.
When the measurement arrives, the prediction is updated.

Question 2 **(0.5 point)**

Explain under which assumptions the Kalman Filter is the optimal solution for a tracking problem. Give an example in which these assumptions hold.
Kalman Filter is the optimal solution for the tracking problem under the LDM (Linear Dynamics Model) assumptions.

Question 3 **(1 point)**

Traffic cameras provide real-time information of traffic at different areas. Consider an application in which traffic is analysed using car trajectories. These trajectories are obtained tracking cars from fixed traffic cameras. Consider the images provided by the traffic cameras showed below:

a) Discuss, for each camera, the convenience of using a Kalman filter or Particle filters to track vehicles.
Based on the dynamics of vehicles that are supposed to follow the road, Kalman filter may be suitable to track cars from Camera 2, while Particle filter is more convenient for Cameras 1,3.

b) The company responsible of the traffic cameras has available a Particle filters implementation. Which is the objective function that should be estimated? Give the mathematical expression of the approximation provided by Particle filters.
Particle filter expression

c) Suppose that the state of a vehicle is defined using its bounding box and its velocity. Write the state vector.
$x = [x\ y\ h\ w\ v_x\ v_y]$

d) Explain the concept of degeneracy. Can it be avoided? Assume that we have five particles as input of the Resampling step with weights:

$w_1 = 0.42$ $w_2 = 0.01$ $w_3 = 0.13$ $w_4 = 0.37$ $w_5 = 0.07$

Explain what is the expected result after this step. If this process is repeated more than one time with the same input, is the same output obtained?

Degeneracy: all the weight associated with particles is concentrated in only one particle. It cannot be avoided, but there are some techniques to minimize its effects.
After the resampling step, particles with small weights are replaced with particles with higher weights statistically. This means that the result could be different each time the process is performed.

e) What is the number of estimates processed by the Kalman filter and Particle filters? What approach may be more useful to track cars that can be eventually occluded? (Camera 3)
Kalman filter processes one estimate while particle filter processes a number of estimates equal to the number of particles. Particle filters are more robust to occlusions because when the object is dissocluded any estimate (particle) can be used to recover the tracked object.

Part IV Constantine Butakoff (1 point)

Question 1 **(0.5 point)**

Given a large set of images of faces and nothing else, describe the steps required to build the Point Distribution Model (PDM) of the faces (starting with the landmarking and ending with the shape model formulas, do not describe the algorithm for matching the model to an image, also provide the steps of the PCA).

Answer (one of the possibilities):

1. Represent the faces by a set of landmarks placed along the facial contours on all the images
2. Align the shapes using the Procrustes analysis
3. Represent every shape by a vector of concatenated coordinates
4. Calculate the covariance matrix C of the shape vectors, do the eignedecomposition: C = VDV'
5. Every shape X can be represented by $X = \bar{X} + Vb$, where b is a vector of parameters and $\bar{X}$ is the average shape.

Question 2 **(0.5 point)**

Describe the necessary conditions for the landmark location (during landmarking of large datasets for PDM construction) and the assumptions on the distribution of the shape vectors in the classical ASM and AAM.

Answer:

1. The landmarks with the same index, in all the images have to represent the same feature or be in the same anatomical position (in the example of the faces). I.e. if the 1st landmark is in the left corner of the left eye, then it has to be in the left corner of the left eye in all the images.
2. ASM and AAM assume that the shape vectors have normal distribution. The algorithms actually estimate the parameters of this distribution.

Part V Josep Ramon Casas (1 point)

Question 1: Model-based tracking (Mocap) **(0,5 point)**

Motion capture (Mocap) can be used to transfer the motion captured from a human being and animate with this an avatar.

1) What is marker-based Mocap? Where are the markers located? List possible tools/devices for marker-based Mocap.
2) Does any method of marker-based Mocap require image detection and tracking? If so, mention one case.
3) In AV production, the captured motion may be used to animate either a human avatar or a non-human character. Discuss the challenges related to the body models used for capture and animation/rendering when transferring the motion captured from a person to a non-human avatar.

Answer

1) Marker-based Mocap consists in recording human movement of a given set of markers located on a performer. Markers are located on articulations (neck, hip, shoulders, elbows, knees…) and end effectors (hands, feet and head). Optical, magnetic or mechanical devices
2) Yes: optical markers should be detected and tracked from the images recorded with a camera
3) The articulations defined on the human body model should be mapped onto the character body model. Such mapping should relate end effectors from both models and articulation points from both models. Some control points may be either not mapped or interpolated from the existing capture information

Question 2: Model-based tracking (Pose inference) **(0,5 point)**

Markerless Mocap should infer the body pose and track the articulations along time.

1) What is "generative" model-based tracking? What are the phases of modeling and estimation?

2) Human body pose and tracking requires an estimator that works by comparing the observations to the model. Why is this task challenging when the human body model is an articulated body model represented as a graph or skeleton?
3) What is understood by the "dimensionality problem" in model-based tracking? How do computer vision algorithms face this problem?

Answer

1) Generative model-based tracking is an analysis-by-synthesis using a Human Body Model (HBM). The modelling phase involves the HBM definition, likelihood and matching functions. The estimation phase aims to find the most likely pose according to the observations
2) Observations are 2D images under camera perspective from which features have to be extracted to detect articulations and relate them to the HBM. When the articulated model is represented as a skeleton graph, the comparison of different pose hypothesis of the model with the observations requires a fleshing out or rendering from the skeleton parameters.
3) The estimation process works by generating pose hypothesis for the articulation parameters. The degrees of freedom (DoF) of the model makes almost impossible to evaluate all the possible hypothesis. This is known as the dimensionality problem. A dimensionality reduction is needed to limit the number of evaluated hypothesis as an optimized (non-full) search to find the most likely pose.

Part VI Javier Ruiz (1 point)

Question 1 **(0.5 point)**

Describe briefly the different types of marker/suit based capture systems that can be used for gesture and activity recognition. Enumerate their advantages or disadvantages with respect to marker-less based systems.

Answer: lesson 8, slides 11, 20

Question 2 **(0.5 point)**

Explain the advantages of using a mixture of experts in gesture classification systems. Enumerate at least three classification techniques that can be used in gesture and activity recognition systems.

Answer: lesson 8, slides 54, 60, 61

Part VII Jordi González (1 point)

Question 1 **(1 point)**

Imagine that the Barcelona City Council asks you to deploy a computer vision system to analyze the behaviors of cars in the main highways of the city (the car tracking module is already provided), in order to make street modifications (driving directions, allowable turns, etc).
Which family of behavior modeling would you choose, top-down or bottom-up? justify the selection, list its main characteristics, describe a proper representation to model car behaviors, and explain which limitations would have your solution.

Answer:
The solution would be a bottom-up behavior model, since the goal is to analyse the trajectories of the cars over time.

The main characteristics of a bottom-up behavior model are:

- A set of normal behavior patterns is learnt
- The model can change over time if new patterns are observed
- The model can be generically applied in any street
- Robust to noise (if considered in the learning step)
- The behavior model is not specified beforehand
- There is no need of an expert for embedding semantic knowledge

A proper representation of car trajectories would be a Mixture of Gaussians, each Gaussian centered on a particular street position. Another one would be a spline, in which the control points would be the tracked points of a car.

The main limitations would be:

- Very limited semantic explanation can be extracted
- Difficult to interpret by users

Part VIII Xavier Giró (1 point)

Question 1 **(0.5 point)**

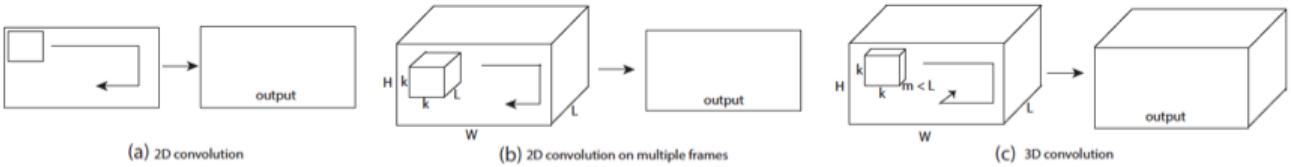
Which are the three basic approaches by which video data can be processed with convolutional neural networks ?

Answer:

(a) Independent 2D convolutions for each single frame.

(b) 2D convolutions which are fused at some stage of the convnet (early, late or slow fusion).

(c) 3D convolutions over a volume of spatio-temporal data.



(a) 2D convolution (b) 2D convolution on multiple frames (c) 3D convolution

Question 2 **(0.5 point)**

Name five software environments which allow developing and testing convolutional neural networks.

Answer:

Caffe, Theano, Torch (Overfeat), TensorFlow, MatConvnet and Computational NeTwork Kit (CNTK).