Master in Computer Vision Barcelona

UAB UOC UPC upf

Module: **M4. Video analysis** **Final exam**
Date: February 23rd, 2017 **Time: 2h30**
Teachers: Montse Pardàs, Ramon Morros, 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.
- All results should be demonstrated or justified.

Question 1: **0,5 Points**
Name two features that can be used for video shot segmentation. Describe the algorithm that you could use to detect shots with these features. In which situations would you use each one of these features? In which cases would the both fail?
Features: frame difference, frame histogram comparison, displaced frame difference
Algorithm: Compute this feature for each frame and binarize, or marker detection and region growing
We coud use FD when there is not much motion within a shot, and DFD when there is more motion
They would fail for instance in subtle shot transitions or when the internal transitions within a shot are too high

Question 2: **0,5 Points**
Describe the main ideas of foreground segmentation based on statistical pixel modeling. Describe similarities and difference of modeling the pixel with a Gaussian Mixture Model within the Stauffer and Grimson approach and with Kernel Density Estimation.
Estimate a statistical model such as a Gaussian or a MoG for each pixel. Assign a pixel to bg if it can be explained by its corresponding model and to fg if it cannot.
Both are modeling the appearance of pixels with a statistical model. Stauffer and Grimson with a MoG and KDE with a non parametric model: the histogram of the n most recent pixel values smoothed with a Gaussian kernel.

Question 3: **0,5 Points**
Explain why shadows can be a problem in a foreground segmentation method based on statistical pixel modeling. Propose a system to try to solve this problem.
Because the appearance of a shadow pixel does not fit with the background model, and thus they are detected as foreground. Moreover, they move as the object casting them. A postprocessing can be done using thresholds on the Color and Brightness distortion to remove pixels detected as foreground that correspond to shadows.

Question 4: **0,5 Points**
a) Explain how the motion information can be used to perform video segmentation.
b) Explain how image segmentation systems can be extended to video segmentation. What would be the main problem of this approach?
a) For instance in a bottom-up approach, to merge regions with similar motion. Or by using the optical flow to compute the neighborhood of a pixel in the next image.
b) For instance, defining the neighborhood of a pixel in 3D (2 spatial dimensions + time). If there is large motion then pixels belonging to the same object in different frames can become disconnected.

Question 5: **1 Point**

a) Starting from the brightness-constancy equation (1) derive the Lucas-Kanade equation for the optical flow (2), explaining the necessary assumptions made.

$$I(x,y,t)=I(x+\Delta x,y+\Delta y,t+\Delta t) \quad (1)$$

$$\nabla I\cdot V^{T}+I_{t}=0 \quad (2)$$

b) To solve the aperture problem, equation (2) is applied applied to a region around the given pixel, assuming coherent motion inside this region. In this case, the equation is expressed as A·V = b, where matrix A contains information about the spatial gradients and vector b about the temporal gradients in this region. Explain under which conditions this problem has a solution.

a) See slides, MCV_M4_L3_Motion, pp 52
b) See slides MCV_M4_L3_Motion, pp 62

Question 6: **1 Point**

a) Explain **concisely** (maximum 10 lines, no need to use equations) the Kalman filter approach for object tracking, along with the main assumptions
b) We want to track the ball in a tennis match. Discuss the appropriateness of the Kalman method for this application
c) The Kalman filter's goal is to recover the most likely state of an object given a set of past and current measurements of the object's state and knowledge of the dynamics model that the object follows. It is a two step process that combines a prediction of the object state based on a dynamics model (propagates state pdf forward in time) and a correctionusing Bayes theorem to modify prediction pdf based on current measurement. To compute probabilites using this Bayesian approach, it is assumed that the process is Markovian (Current state X_t depends only on previous state X_{t-1}). Independence (measurement depends only on the current state) is also assumed. When both **dynamics and measurement models are linear** and subject to Gaussian noise, the Kalman filter provides the optimal solution to the estimation problem.

d) Kalman requires the dynamics model to be linear. In a tennis match, when the ball hots the racket or the floor, it changes it's direction in a non-linear way. For this reason, Kalman would not be optimal in this case.

Question 7: **1 Point**

a) In a Particle Filter framework, explain degeneracy phenomenon, how it is measured and one practical method to reduce its effects.
b) In a SIR Particle filter tracking application, describe how an estimate of the position of the tracked object be obtained.

a) Lecture 5 slides pp. 30
b) Computing the mean state of the system or taking the position of the particle with higher weight after normalizing the particle weights

$$E[X_k] = \sum_{n=1}^{N_s} w_k^i x_k^i$$

Question 8: **0,5 Points**

1) What are the challenges of vision-based HMA?
2) Cite at least one use case of HMA in the industrial, entertainment, medical, video analytics and computing domains
3) What is Mocap? How is Mocap used in AV production?
4) Vision-based Mocap may be related to a simple tracking problem or to a more complex pose inference problem

1) Challenges: perspective, non-rigid object type, scene clutter, imprecise semantics of actions and human motion
2) Industrial: ergonomics, fashion, movie animation; Entertainment: gaming, consoles, edutainment; Medical research: bio-mechanics, surgery, sports; Video analytics: pose, gesture, gait, behavior; Computing: HCI, emotion-aware
3) Recording human movement and translating it onto a digital model. In AV production, it is used to animate virtual models
4) Mocap can be implemented by tracking visual markers located on the human body (marker-based) or by inferring the pose of a (hidden) skeleton model from a visual set of observations (markerless)

Question 9: **0,5 Points**

Pose inference is a baseline for HMA. It is tackled with discriminative and generative approaches.

1) Explain how discriminative approaches are posed as a learning/classification problem and whether the use of a Human Body Model (HBM) is required in this case.

2) What about generative pose inference approaches? Do they require a human body model? Why? What are the two phases of generative approaches and what steps do they involve?

1) Discriminative approaches just learn a mapping from visual observations to articulated body configurations, and then perform classification with the trained classifier. The learnt configurations can be just labels for a set of training data, not necessarily related to an explicit HBM.

2) Generative approaches perform analysis-by-synthesis necessarily using some kind of explicit 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

Question 10: **0,5 Points**

List briefly three different applications where gesture recognition can be used.

Answer: lesson 8, slide 8. Human-Computer Interaction (HCI). Deaf people assistance. Synthesis and Animation (films, computer games). Surgery / Medical applications, Virtual reality, etc.

Question 11: **0,5 Points**

Explain at least one advantage and disadvantage of using deep learning techniques to perform gesture and activity recognition in video sequences.

Advantage: Better classification performance. Disadvantage: Large training dataset needed.

Question 12: **1 Point**

The Barcelona City Council is getting worried about the exponential increase of cyclists in Barcelona, since their behavior is sometimes very chaotic and anarchic, even dangerous for senior citizens (there have been several casualties among older people indeed).

Imagine that your own spin-off receives an offer from the Barcelona City Council:

- to semantically analyze the interactions between cars, pedestrians and cyclists in 1,000 pedestrian crossings;
- to detect collisions, overtakes and chasings, as defined by an expert; and
- to send the semantics of such detected incidents to (two or three) human operators, without having to monitor 1,000 screens at the same time.

Which family of behavior modeling would you choose (top-down or bottom-up)? (i) justify the selection, (ii) list its main characteristics and benefits within this problem, (iii) describe the steps required to define a proper representation to model semantic interactions in pedestrian crossings, and (iv) explain which limitations would have your solution.

(i) The solution would be a top-down behavior model, since the goal is to analyze the semantics of the interactions between different agents in a pedestrian crossing over time.

(ii) Top-down behavior models can be specified beforehand:

- All the knowledge can be provided by an expert
- Can provide accurate semantic descriptions
- Easy to understand by end users

(iii)

- Step 1: Modeling scenes: To associate semantic tags to the ground plane to better analyze different pedestrian/car/cyclist behaviors in a pedestrian crossing scenario.
- Step 2: Modeling behaviors: By using semantic models like Petri Nets, Graph Structures or Semantic Tree structures, which allows to incorporate specific behaviors involving different types agents and to reason from tracking data to generate predicates with different semantics.
- Step 3: Natural-Language generation: To report to the end-users using texts (and images) only when the pre-defined behaviors defined by the experts are detected.

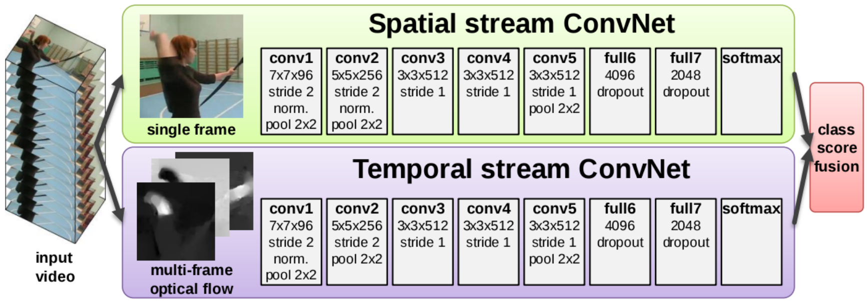
(iii) Limitations:

- Not robust to noisy observations
- Useful for restricted environments.
- Not evolvable

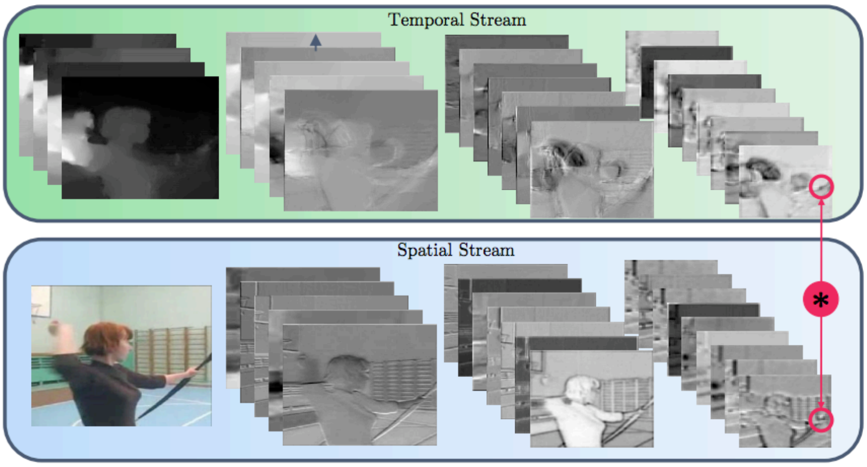
Question 13: **0,5 Points**

Draw a scheme of the two-stream deep convolutional network used for action recognition in videos, as presented by Symonian and Zisserman (NIPS 2014) and Feichtenhofer, Pinz and Zisserman (CVPR 2016). Indicate the main difference between the two proposals.

Symonian and Zisserman (NIPS 2014)



Feichtenhofer, Pinz and Zisserman (CVPR 2016)

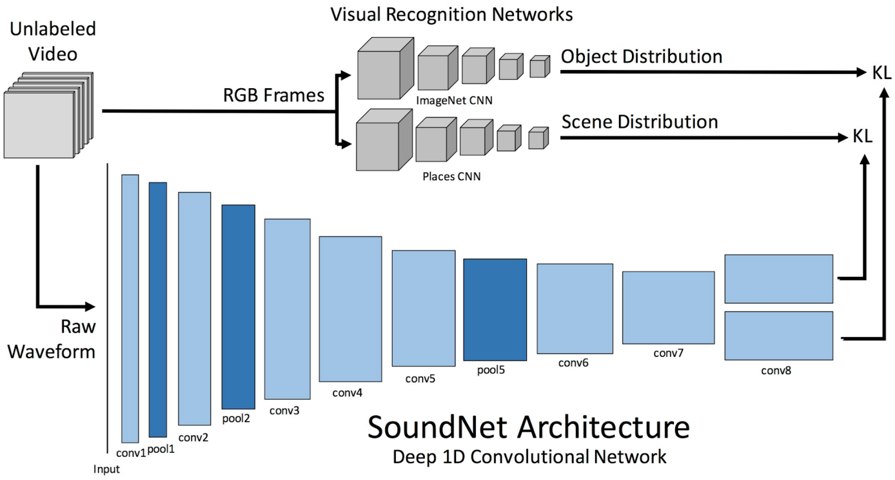


The main difference between the two solutions is how the spatial and temporal streams were fused. In NIPS 2014 the final class scores where fused, while in CVPR 2016 the fusion was performed at the last convolutional layer.

Question 14: **0,5 Points**

The SoundNet architecture is a deep 1D convolutional network that learns sound representations from videos. These audio features are learned from labels extracted from the visual information of the video. How are these labels obtained ?

These labels are obtained by extracting the frames of the video and analysing them with two different convolutional neural networks: one trained on the ImageNet dataset and another one trained with the Places dataset. This way, weak labels are generated to represent the objects and locations depicted in the videos.



Question 15: **0,5 Points**

The Long Short-Term Memory units contain different types of gates, under the form of a sigmoid unit that controls the input and output information. Which are the name of these units ?

Forget gate, input gate and output gate.

Question 16: **0,5 Points**

What are the types of the input and output data of the convolutional neural network used in *Vid2Speech* by Ephrat and Peleg (ICASSP 2017) ?

The input data are video frames, and the output data are LSP audio feature vectors. This is not an end to end solution, so no audio (speech) samples are generated at the output.