Master in Computer Vision Barcelona

UAB UOC UPC upf.

Module: **M6. Video analysis** **Final exam**
Date: April 26, 2018 **Time: 2h30**
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.

Part 1

Question 1: **0,5 Point**

a) Describe three features that can be used to produce shot segmentation.
b) Describe an algorithm to perform shot segmentation using one of these features.
c) In case of strong motion, which features would you use? And in case of a gradual transition between shots?

a) A temporal gradient can be computed with one of these features: Frame difference (FD), Frame histogram comparison, Displaced frame difference (DFD

$$FD = \sum_{\vec{r}} \left(I(\vec{r},t) - I(\vec{r},t-\Delta t) \right)$$

$$DFD = \sum_{\vec{r}} DFD\left(\vec{r}, \hat{D}(\vec{r})\right) = \sum_{\vec{r}} \left(I(\vec{r},t) - I(\vec{r} - \hat{D}(\vec{r}),t-\Delta t) \right)$$

b) This feature is computed for each frame, and a threshold is applied. We could also apply a segmentation algorithm to the resulting 1D function in order to find homogeneous regions and the transition between these homogeneous regions. A prefiltering of the 1D function would improve the results.
c) For strong motion, DFD would be more appropriate. Histogram comparison would also work if the same objects remained on the scene. In case of gradual transition FD would be better, but we would require a low threshold, which would produce a high number of false positives. In this case, an algorithm to find homogenous regions would work better.

Question 2: **1 Point**

a) Assume you want to detect the foreground in a scene by background subtraction. Explain how you would build a single gaussian background model if you do not have a training sequence without foreground objects. Explain how to initialize the model and how to update it.
b) Write an algorithm which applies the previous model to detect foreground, shadow and highlight pixels.

a) We need to model one Gaussian for each pixel. We could use the median value of the first N frames for modeling the mean of the Gaussians. This would make the algorithm more robust to the appearance of foreground objects in these frames. The variance can be estimated with this initial sequence, but we would obtain a too large value for pixels with foreground objects. Thus, we should establish a maximum value for the variance.
The mean and the variance could be updated at each frame according to:

$$\mu_{(i,j)}(t) = \begin{cases} (1-\rho)\mu_{(i,j)}(t-1) + \rho I(i,j,t), & \text{if } (i,j) \in \text{background} \\ \mu_{(i,j)}(t-1), & \text{otherwise} \end{cases}$$

$$\sigma^2{}_{(i,j)}(t) = \begin{cases} (1-\rho)\sigma^2{}_{(i,j)}(t-1) + \rho \left(I(i,j,t) - \mu_{(i,j)}(t) \right)^2, & \text{if } (i,j) \in \text{background} \\ \sigma^2{}_{(i,j)}(t-1), & \text{otherwise} \end{cases}$$

b) First we would use a decision rule such as

$$\text{If } \left| I(i,j,t) - \mu_{(i,j)}(t-1) \right| < \alpha \sigma_{(i,j)}(t), \text{ then } (i,j) \in \text{background}$$

For the pixels detected as foreground we could compute Colour Distorsion and Brightness Distorsion and apply the following rules:
IF CD < 10 THEN
IF 1 > BD > 0.5 -> SHADOW
IF 1.25 > BD > 1 -> HIGHLIGHTING

ELSE FOREGROUND

Question 3: **0,5 Point**

Explain the principles in which background subtraction using Eigenbackgrounds is based. Describe the algorithm.
See slides 68 and 69 of 6.2.

Question 4: **1 Point**

We wish to build an Active Shape Model to find the contours of flat hands facing up (see the example picture).

a) Propose a suitable set of landmarks
b) A database of 100 images with different hand positions is manually labeled, obtaining the vectors y_i, (1<=i<=100). Which should be the components of these vectors?
c) These vectors are first aligned using Procrustes analysis. How would you align a given vector y_k to a defined mean vector $\bar{y}$?
d) Once the vectors have been aligned (z_i), the mean vector is subtracted ($\bar{z}$) and the covariance matrix of this data is computed. Explain what you need to do next to be able to define a hand shape with a reduced number of parameters. How could you choose this number of parameters?
e) If you have a new vector of data, how could you determine if it corresponds to a hand shape?
f) Describe the main steps of the Active Shape Model algorithm to find the position of a hand in a new image.

a) Landmarks should be in well defined corners (points of high curvature). Additional points can be added along the boundary, for instance at intermediate points.
b) If n is the number of landmarks used, each vector would have 2n elements, corresponding to the x and y coordinates of the landmarks.
c) We should minimize E, as expressed in slide 17 of 6.8.
d) We need to do Eigenvector decomposition of the Covariance Matrix. Then, dimensionality reduction is done removing the eigenvectors corresponding to the smaller eigenvalues. To decide the number of eigenvectors that we need to keep we can put a threshold on the global variance, which is the addition of the eigenvalues, or use the elbow method (see slide 19 of 6.8)
e) To analyze the plausibility of a shape we first aling to the mean shape, then substract the mean shape and project into the eigenspace, obtaining the parameters b. Then, we have to check if, for each of the eigenvalues corresponding to the kept eigenvectors:

$$|b_j| < \beta \sqrt{\lambda_j}$$

f) 1) Initialize in a roughly good position
Iterate:
2) Displace every vertex to minimize Mahalanobis distance to the mean profile
3) Enforce plausability

Part 2

Question 5: **1 Point**

Block matching:

a) Explain what a motion-compensated image is.
b) Explain the difference between forward and backward motion compensation. Which version is used in video coding? Why?

a) The reference image to which the optical flow has been applied is known as **motion compensated image**.
b) Slides 31-34. Slide 40. Backward compensation is used as the compensated images has no 'holes' and has lower error estimation energy.

Question 6: **1 Point**

Optical flow:

a) Explain the advantages and disadvantages of the Horn-Schunk method over local methods such as Lucas-Kanade for optical flow estimation.
b) The energy functions for the Horn-Schunk and Brox [2004] methods are given by equations [1] and [2] respectively. Describe the terms of the energy functions in both cases and explain the main differences between the two approaches.

$$E = \iint (I_x u - I_y v + I_t)^2 + \alpha(|\nabla u|^2 + |\nabla v|^2) dxdy \quad [1]$$

$$E = \iint \Psi(|I(x+w) - I(x)|^2 + \gamma|\nabla I(x+w) - \nabla I(x)|^2) + \alpha\Psi(|\nabla_3 u|^2 + |\nabla_3 v|^2) dxdy \quad [2]$$

a) HS yields a high density of flow vectors, i.e. the flow information missing in inner parts of homogeneous objects is *filled in* from the motion boundaries. LK can only compute reliable vectors for corner-like points. However, HS is more sensitive to noise tan LK.

b) HS is a non-robust method, Brox uses robust estimation. HS uses a linearized version of the brightness-constancy equation, Brox uses the non-linearized version. Brox uses a multi-resolution approach versus the single resolution approach of HS.

Question 7: **1 Point**

The Kalman filter equations for the case of 1D tracking are given by:

$$\mu_k^- = d \cdot \mu_{k-1}^+ \qquad (\sigma_k^-)^2 = \sigma_d^2 + (d \cdot \sigma_{k-1}^+)^2$$

$$\mu_k^+ = \frac{\mu_k^- \cdot \sigma_m^2 + m \cdot z_k \cdot (\sigma_k^-)^2}{\sigma_m^2 + m^2 \cdot (\sigma_k^-)^2} \qquad (\sigma_t^+)^2 = \frac{\sigma_m^2 \cdot (\sigma_k^-)^2}{\sigma_m^2 + m^2 \cdot (\sigma_k^-)^2}$$

a) Describe how the filter behaves if there is no prediction uncertainty. Same if there is no measurement uncertainty.

b) Suppose that we want to track an object whose dynamics model is described by:

$$\Phi(x_{k-1}, v_{k-1}) = Dx_{k-1}^2 + \Sigma_d$$

Explain the adequation of the Kalman filter to solve the tracking problem in terms of optimality of the solution.

a) If there is no prediction uncertainty, measurement is ignored. If there is no measurement uncertainty, prediction is ignored

b) The dynamics model is not linear as it depends quadratically of x. In this case the Kalman filter solution is not optimal and may fail to track the object.

Question 8: **0,5 Point**

Particle filters:

When using Importance Sampling, the weights w^i of the particles are given by eq. [3]. However, in many cases, the distributions p(x) and q(x) can be evaluated only up to a normalizing constant (equations [4-5]). Describe how the normalized weights w^i can be computed from the un-normalized distributions $\tilde{p}(x), \tilde{q}(x)$

$$w^i = \frac{p(x^i)}{q(x^i)} \quad [3] \qquad q(x) = \frac{\tilde{q}(x)}{z_q} \quad [4] \qquad p(x) = \frac{\tilde{p}(x)}{z_p} \quad [5]$$

The non-normalized can be computed using . Then, the weights can be normalized by dividing by the sum of all the weights.

Part 3

Question 9: **0,5 Point**

1) What is Mocap? Why may we need Mocap for computer vision?
2) What is "generative" model-based tracking? Describe the phases of modeling and estimation
3) Can you tell what are “Deformable Part Models” (DPM), and how can they be applied to model-based tracking?

1) Mocap is recording human body movement and translating it onto a digital model.
Marker-based Mocap may be used as ground truth to train or evaluate computer vision-based human body tracking or vision-based Human-Computer Interfaces

2) 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

3) DPMs are based on modeling objects as a set of parts constrained in a spatial arrangement.
A discriminative model captures the appearance of parts providing a response map over an image for each part. Then, in order to consider only feasible configurations of the parts, a global shape deformation model (e.g. a spring model) computes deviations from expected spatial arrangements. Therefore, the discriminative model maximizes local responses

whereas the deformation model minimizes deviations from expected configurations.
In the case of model-based tracking, one may apply several parts-based classifiers at different scales (window sizes), and a Human Body Model plays the role of the global shape deformation model.

Question 10: **0,5 Point**

List briefly at least two advantages and disadvantages of using body models for pose and gesture recognition systems.

Advantages: Easier to generalize, able to put priors (physical constraints), independent of marker placement / feature selection

Disadvantages: Initialization is usually needed, need of tracking, prone to lose track

Question 11: **0,5 Point**

Explain at least one advantage and disadvantage of using deep learning techniques to perform pose and gesture recognition in video sequences. List briefly at least two deep learning architectures that can be used for **dynamic** gesture recognition.

Advantage: Better classification performance.

Disadvantage: Large training dataset needed.

3D-CNN, RNN, LSTM, combination of image + motion features

Part 4

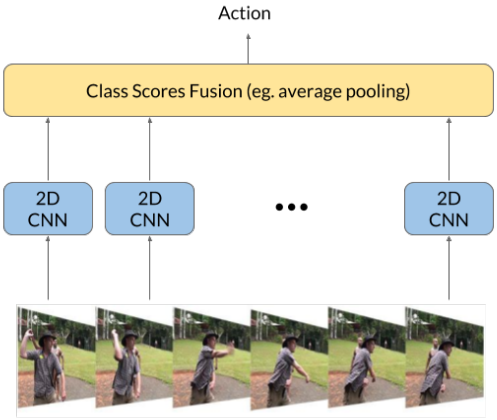
Question 12: **0,5 Point**

Consider a deep neural network able to solve the action recognition task in a video clip composed of K=6 frames. Draw a scheme depicting a basic architecture for each of the following set ups:

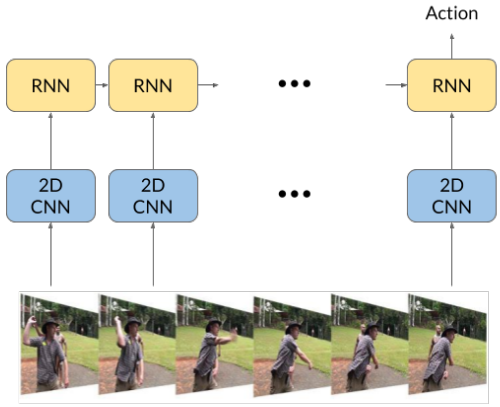
a) Single frame model
b) CNN + RNN
c) 3D CNN
d) Two-Stream

Solution:

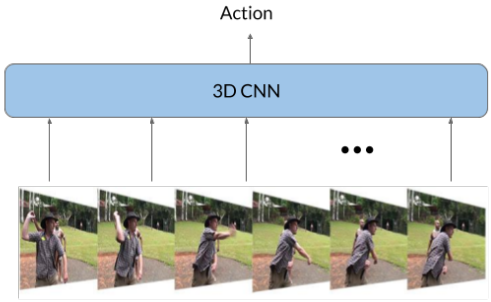
a) Single frame model



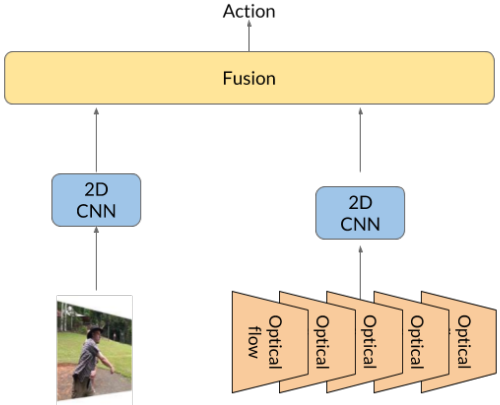
b) CNN+RNN



c) 3D CNN



d) Two-stream



Question 13: **0,5 Point**

Enumerate eight tasks that can be formulated over a video to learn visual features following a self-supervised approach. Organize your answer based on the modality of the supervisory signal (vision and audio).

Vision

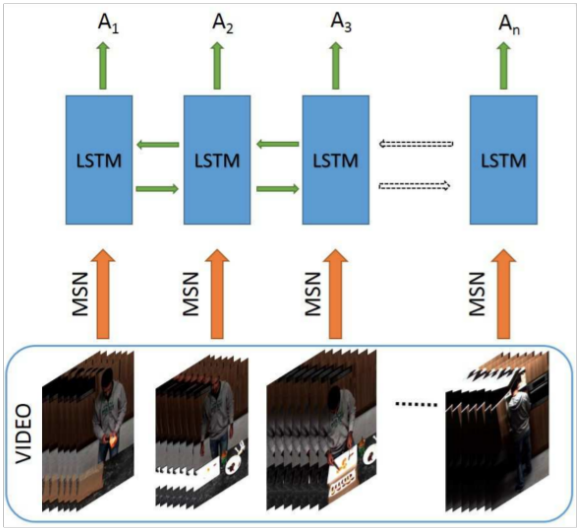
- Frame reconstruction (auto-encoding).
- Frame prediction
- Patch prediction.
- Binary classification of correct/incorrect order of frames.
- Prediction of a temporally odd frame.
- Segmentation based on motion (eg. optical flow)

Audio

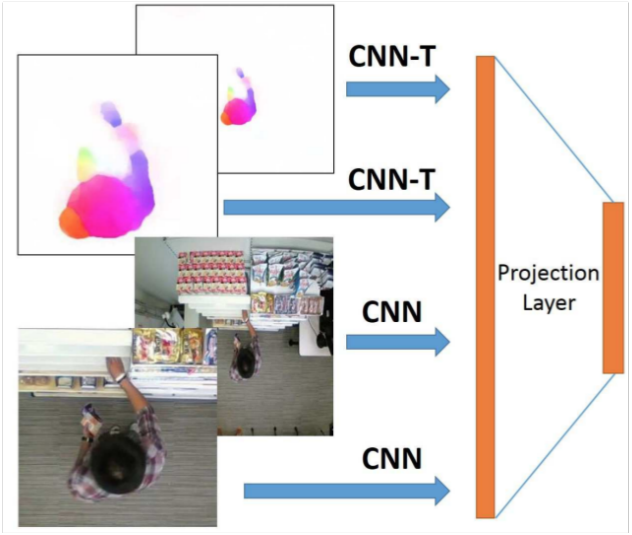
- Prediction of audio features (eg. Cochelogram).
- Correspondence between pairs of audio and visual frames.

Question 14: **0,5 Point**

These two figures are extracted from Singh et al, CVPR 2016.



a) Overall system



b) Multi-Stream Network (MSN)

Answer the following questions:

a) Which video analysis task does this model solve?
b) Explain figure a) in less than 100 words.
c) Explain figure b) in less than 100 words.

a) Action/Activity Detection
b) Short chunks of a video are given to a multi-stream network (MSN) to create a representation for each chunk. The sequence of these representations is then given to a bi-directional LSTM, which is used to predict the action label, Ai.
c) The multi-stream network uses two different streams of information (motion and appearance) for each of two different spatial croppings (full-frame and person-centric) to analyze short chunks of video. One network (CNN-T) computes features on pixel trajectories (motion), while the other (CNN) computes features on RGB channels (appearance).

Question 15: **0,5 Point**

The Connectionist Temporal Classification (CTC) loss, used in LipNet for maching audio and vision with language, avoids the need of alignment between input and output sequences by:

a) Adding an additional token ("word") at the output classifier. How is this token referred to?

b) Removing two patterns from the predicted sequence. Which are these two patterns?

a) A blank “_” token.

b) Blank tokens and repeated words are removed.