Computer Vision (INFOMCV) - Exam

2016-2017, Utrecht University

April 10, 2017
Duration: 13.30 - 16.30

Instructions:

1. *Write your name and student number on every separate answer sheet.*
2. *The total number of points is 100.*
3. *There are always potentially multiple correct answers in the multiple-choice questions. Missing one correct answer or providing an incorrect answer will cost you half of the points.*
4. *Write your answers as complete as possible. However, adding irrelevant information might decrease your score. Examples and drawings cannot replace the textual answer. You can use the final page if you need more space.*
5. *Ensure that your handwriting is readable. You can answer in English or Dutch.*
6. *You are not allowed to speak with other students, use your phone or additional materials.*
7. *You are allowed to leave the room anytime after 14.00, by first handing in all answer sheets. Show your ID when handing in your work.*

Good luck!

1. **Image formation (4 points)**
   What is the unit of focal length f?

   (a) meter

   (b) degree or radial, same convention as used in rotation part of the extrinsic camera matrix

   (c) pixel

   (d) no unit, it is just a conversion

   Answer: ____________________

2. **Image formation (10 points)**
   We rotate a camera around the camera center (local world origin). Write down the extrinsic camera matrix $[\mathbf{R} \quad \mathbf{t}]$ and intrinsic camera matrix $\mathbf{K}$, and state explicitly which elements change under the rotation.

3. **3D reconstruction (12 points)**
   We want to make a 3D voxel reconstruction from a sequence of image frames. Typically, we have a sequence of RGB frames, a background frame and a threshold as input. In this case, however, we only have a single RGB starting frame, a background frame, a threshold and a sequence of dense optical flow fields from each image to the next. Can we obtain a 3D voxel model in the final frame? If we can, explain how. You may use pseudocode. If we can't achieve this, explain which information we would need to make it possible.

4. **3D reconstruction (6 points)**
We have a physical setting as in Assignments 2 and 3, with four calibrated cameras aimed at a certain space. In silhouette-based volume reconstruction, when a user wears a t-shirt that has the same color as the background in one of the views, there will be a hole in the voxel model. Explain how we could mitigate this issue.

5. **Clustering (8 points)**
We will construct a color histogram. Instead of using equally-sized bins, we will determine the most common colors using a set of training images. We take the color values of each of the pixels in the training images and apply K-means clustering on this set, with $k = 100$ clusters and Euclidian distance as the distance function. Given a new image, describe how we obtain the feature vector that represents the color distribution. Mention the processing steps and output explicitly. You may use pseudocode.

6. **Image descriptors (4 points)**
Which is **not** and advantage of Canny edge detection?

(a) Thicker edges are reduced to 1-pixel thick edges
(b) It speeds up the edge detection process
(c) Incidental or small traces of edges can be removed
(d) Missing high-contrast pixels are interpolated

Answer: ____________________

7. **Image descriptors (8 points)**
For a found SIFT point, explain in detail how we can determine (1) its orientation and (2) its scale.

8. **Optical flow (4 points)**
Which of the following statements are true?

   (a) Sparse optical flow is calculated at a regular grid
   (b) The main assumption of optical flow is that the sequence of frames are stabilized
   (c) In-plane rotation will cause issues in estimating optical flow
   (d) The use of an image pyramid in Lucas-Kanade is essentially a coarse-to-fine strategy

   Answer: ____________________

9. **Training and testing (13 points)**
We have a training set of n frames taken from soccer matches. In the training set, all soccer players are indicated by (1) the frame in which they occur and (2) a bounding box. In total, there are m soccer players. Assume that we can describe the image bounded by the bounding box as a HOG descriptor. Explain the steps to train a classifier that can classify a selected region, encoded as a HOG descriptor, as a soccer player or "other". Your pipeline should make use of hard negative mining. Be explicit in (1) how you obtain your initial set of negative samples and (2) your augmented set of negative samples. You may use pseudocode. A schematic drawing can be used to illustrate your steps, but does not replace your textual explanation.

10. **Performance measures (6 points)**
    We consider a detection task, in which we offer a trained classification model parts of an image in a sliding window approach. The output of the classification is a binary label indicating the presence of the object of interest. Explain why accuracy is not a suitable performance measure.

11. **CNNs (6 points)**
    We have six 5×5 filters for an input of 32×32 with depth 7. We use a stride of 1 and padding of 1. What will be the output size of the produced activation map?

    Answer: ____________________

12. **CNNs (4 points)**
    The pooling operation has been used to decrease the size of the input to avoid a large computational overhead. Which of the following pitfalls are correct?

    (a) Detailed features are lost
    (b) More parameters
    (c) Bigger kernels
    (d) More convolutions

    Answer: ____________________

13. **CNNs (4 points)**
    In CNNs, which of the following options is likely to reduce the risk of overfitting?

    (a) Adding more layers in the network, effectively increasing the number of parameters.
    (b) Adding more unique training samples
    (c) Duplicating your training data
    (d) Training for an increased number of iterations
    (e) Using a non-convex loss function in combination with a larger batch size

    Answer: ____________________

14. **CNNs for video (11 points)**
We have a set of 1-second (25 frames per second) videos taken from the side of the road. They contain (a) nothing special (just the environment), (b) a car going from left to right or (c) a car going from right to left. We'd like to distinguish between these classes and use a CNN that can deal with videos. We have discussed several potential techniques. Choose one of these techniques and (1) give the name, (2) specify the input (dimensions and what they correspond to) and (3) specify the output (dimensions and what they correspond to).

That's it! :)