4F12 Computer Vision (solutions) - 2018

Q1(a)(i). Smoothing — remove high frequency noise which is amplified by differentiation

— reduce high-spatial frequency to select scale of interest.

(ii). $$S(x,y) = \sum_{u=-n}^{n} \sum_{v=-n}^{n} g_\sigma(u)\, g_\sigma(v)\, I(x-u, y-v)$$

where $$g_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_2}}$$

and is sampled at $\underline{N = 2n+1}$ discrete values

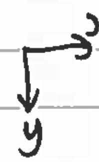
(iii) $$\frac{\partial S}{\partial x} \approx \frac{S(x+1, y) - S(x-1, y)}{2}$$

1D convolution masks

$\frac{1}{2}$	0	$-\frac{1}{2}$

$$\frac{\partial S}{\partial y} \approx \frac{S(x, y+1) - S(x, y-1)}{2}$$

$\frac{1}{2}$
0
$-\frac{1}{2}$



Q1(b)

(i) $C(\underline{n}) = \sum w(x) \left(S(\underline{x}+\underline{n}) - S(\underline{x}) \right)^2$

$$\simeq \sum_{W} w(x) \left[\nabla S . \underline{n} \right]^2$$ by Taylor series expansion

$$\simeq \sum_{W} w(\underline{x}) \left(\underline{n}^T \nabla S^T . \nabla S \, \underline{n} \right)$$

$$\simeq \underline{n}^T \begin{bmatrix} \langle S_x S_x \rangle & \langle S_x S_y \rangle \\ \langle S_y S_x \rangle & \langle S_y S_y \rangle \end{bmatrix} \underline{n}$$

$$\underline{C(\underline{n}) \simeq \underline{n}^T A \underline{n}}$$

(ii) Smoothed values are obtained by convolution with a 2D Gaussian

$$G_\sigma(x,y) = \frac{1}{\sigma_I^2 2\pi^2} e^{-\frac{(x^2+y^2)}{2\sigma_I^2}}$$

where σ_I = "window" $> \sigma_D$ (used for smoothing for differenti

(iii) To maximise/minimize $C(\underline{n})$ need large λ_1, λ_2

$$\lambda_1 \leqslant \frac{\underline{n}^T A^T \underline{n}}{\underline{n}^T \underline{n}} \leqslant \lambda_2$$

$\det A = (\lambda_1 \lambda_2)$ and $\text{trace} A = (\lambda_1 + \lambda_2)$

(iii) Harris corner algorithm

Compute $A = \begin{bmatrix} \langle \ \rangle & \langle \ \rangle \\ \langle \ \rangle & \langle \ \rangle \end{bmatrix}$

Determine $R = \det(A) - \kappa\,(\text{Trace}(A))^2$ $\kappa \simeq 0.04 - 0.06$

Threshold R, and non-maximum suppression

Q2.

(a) (i). Pin-hole camera; central planar projection with no non-linear distortion (ie. line projects to a line)

(b) $$u = \frac{p_{11}X + p_{12}Y + p_{13}Z + p_{14}}{p_{31}X + p_{32}Y + p_{33}Z + p_{34}}$$

$$\therefore V = \frac{p_{21}X + p_{22}Y + p_{23}Z + p_{24}}{p_{31}X + p_{32}Y + p_{33}Z + p_{34}}$$

For known (u, v) and p_{jk} then.

Plane 1 $$X(p_{11} - u\,p_{31}) + Y(p_{12} - u\,p_{32}) + Z(p_{13} - p_{33}u) + (p_{14} - p_{34}u) = 0$$

Plane 2 $$X(p_{21} - v\,p_{31}) + Y(p_{22} - v\,p_{32}) + Z(p_{23} - v\,p_{33}) + p_{24} - v\,p_{34} = 0$$

These 2 planes are NOT parallel and hence intersect to define a ray.

2(c)(i) For unknown p_{jk} and known (u_i, v_i) and (X_i, Y_i, Z_i)
these 2 equations become linear in p_{jk}.

— $\underset{2n\times 12}{A}\ \underset{12\times 1}{\underline{p}} = 0$

— Need $n \geq 6$ points to solve (Use RANSAC to remove outliers)

— Ensure points span 3D volume and are not coplanar.

— Linear equation $\lambda_1 \leq \dfrac{\underline{p}^T A^T A \underline{p}}{p^T p} \leq \lambda_{12}$

(ii) Find $\underline{p}$ corresponding to smallest S.V or E.V of $A^T A$.

— Use this to minimise projection error (non-linear optimisation)

$$\min_{\underline{p}} \sum_{i=1}^{n} \left(u_i - \hat{u}_i\right)^2 - \left(v_i - \hat{v}_i\right)^2$$

where $\hat{u}_i = \text{Projection}\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ and $\hat{v}_i = \text{Projection}\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$

$\begin{pmatrix} s\hat{u}_i \\ s\hat{v}_i \\ s \end{pmatrix} = \begin{pmatrix} 3\times 4 \end{pmatrix}\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$.

2(d) let $Z \approx Z_{av}$

(i)

$$su = p_{11}X + p_{12}Y + p_{13}Z + p_{14}$$

$$sv = p_{21}X + p_{22}Y + p_{23}Z + p_{24}$$

$$s = Z_{AV}$$

(ii) $\frac{\Delta Z}{Z} \ll 1$

$\therefore \begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ 0 & 0 & 0 & p_{34} \end{bmatrix} \Rightarrow \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 2\times 4 \end{bmatrix}\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$

Q3 (a) (i). Each correspondence generates 2 equations in 8 unknowns
Need 4 pts

(ii) Correspondences found by

(1) SIFT features + 128D descriptor
(2) NN matching
or Harris corners and cross-correlation

(iii) RANSAC

- Select 4 pts and matches
- compute H
- count inliers
- repeat until maximize inliers

(iv) $A \overset{8\times1}{p} = 0$

Solve by SVD or optimization.

(3)(b).

(i) Epipolar constraint

Consider stereo cameras. Projection of a point at unknown depth must lie on the projection of the ray in other image.

$$\underline{\omega}'^T F \underline{\omega} = 0$$

In the second view $\omega' l' = \omega' . \underline{l}'$

$$\therefore \omega' . (F\underline{\omega}) = 0$$

$$\therefore \underline{l}' = F\underline{\omega}$$

Homogeneous equation of epipolar line.

(ii) $F\underline{e} = 0$ $\det F = 0$ $\operatorname{rank} F = 0$.

In SVD, third singular value $= 0$

(iii)

$$E = K'^T F K. = T_x R. = U \Lambda V^T \text{ by SVD}$$

$$T_x = U \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} U^T \qquad R = U \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} V^T$$

(iv) $P = K[I|0]$ and $P' = K'[R|T]$. Triangulate by solving $X\ Y\ Z$

4 a) The basic building block of a convolutional neural network comprises three stages: a convolutional stage, a non-linear stage, and a pooling stage. Define each stage mathematically and explain the rationale behind their design. [30%]

Answer

Bookwork that should include the following:

Convolutional stage. 2D filtering $a_{i,j} = \sum_{k,l} w_{k,l} Z_{i-k,j-l}$ where w is a 2D filter and $Z_{i,j}$ are (greyscale) image pixels at the ith jth location. Motivated by the translation invariance of images and the need to reduce the number of parameters to be learned so parameters can be tied.

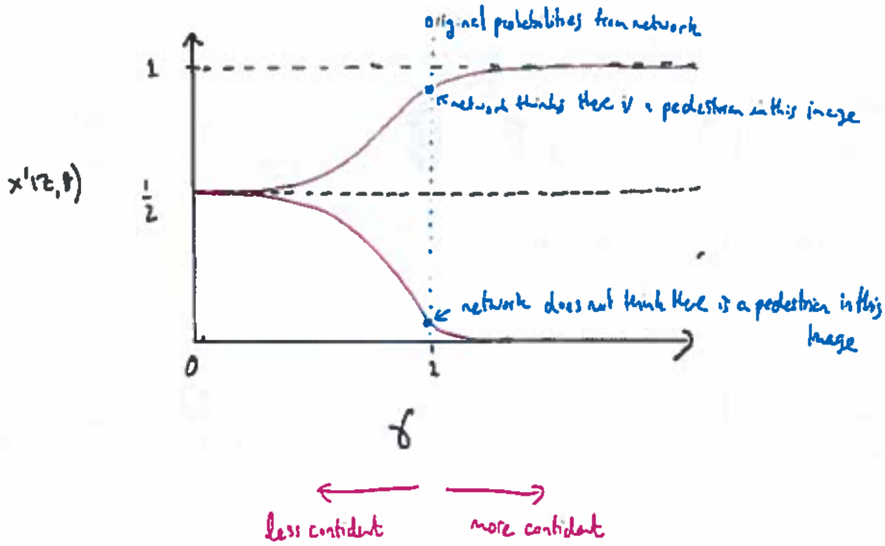
Non-linear stage. A point-wise non-linearity $y_{i,j} = f(a_{i,j})$ where examples include RELU $f(a) = [a]_+$ or sigmoid. Motivated by the fact that non-linear computation needs to be made, and certain point-wise non-linearities followed by linear weighting are universal approximators.

Pooling / sub-sampling stage. Max pooling finds the max value in a region $x_{i,j} = \max_{|k|\leq\tau,|l|\leq\tau} y_{i,j}$ alternatively averaging or subsampling may be used. Motivated by the fact that high-level features are coarser and that subsampling again reduces the number of parameters. It also helps to build in translation invariance.

b) (i) Describe what happens to the output of the new network as γ is swept from zero to infinity and therefore argue how an appropriate setting might improve calibration. [20%]

Answer

The following figure plots the modified network's output as a function of the value of gamma for two example images (one positive prediction and one negative). It shows that a value of $0 < \gamma < 1$ reduces the confidence of the original network's predictions. The sketch was not required and a verbal description would also suffice.



(ii) A new dataset comprising M images $\{Z^{(m)}\}_{m=1}^{M}$ and binary labels $\{t^{(m)}\}_{m=1}^{M}$ will be used to train γ, whilst the original parameters are fixed. Write down the log-likelihood for γ and derive the gradients required for training. [40%]

Answer

Log probability on validation dataset:

$$\boxed{\begin{aligned} \mathcal{L}(\gamma) &= \sum_{m=1}^{M} \left[t^{(m)} \log x_{new}^{(m)}(\gamma) + (1-t^{(m)}) \log (1 - x_{new}^{(m)}(\gamma)) \right] \\ x_{new}^{(m)}(\gamma) &= x'(Z^{(m)}, \theta, \gamma) = \frac{1}{1+e^{-\gamma \underline{w}^{T} \underline{h}(Z,\theta)}} \end{aligned}}$$

Derivative:

$$\frac{d\mathcal{L}(\gamma)}{d\gamma} = \sum_{m=1}^{M} \underbrace{\left[\frac{t^{(m)}}{x_{new}^{(m)}(\gamma)} - \frac{(1-t^{(m)})}{1-x_{new}^{(m)}(\gamma)} \right]}_{} \frac{d x_{new}^{(m)}}{d\gamma}$$

$$\frac{t^{(m)} [1 - x_{new}^{(m)}(\gamma)] - (1-t^{(m)}) x_{new}^{(m)}(\gamma)}{x_{new}^{(m)}(\gamma) \, (1 - x_{new}^{(m)}(\gamma))} = \frac{t^{(m)} - x_{new}^{(m)}(\gamma)}{x_{new}^{(m)}(\gamma)(1-x_{new}^{(m)}(\gamma))}$$

$$\begin{aligned} \frac{d x_{new}^{(m)}}{d\gamma} &= (x_{new}^{(m)})^{2} \cdot \underline{w}^{T}\underline{h}(Z^{(m)},\theta) \, e^{-\gamma \underline{w}^{T}\underline{h}(Z,\theta)} \\ &= (x_{new}^{(m)})^{2} \, \underline{w}^{T}\underline{h}(Z^{(m)},\theta) \left(\frac{1}{x_{new}^{(m)}} - 1 \right) \\ &= x_{new}^{(m)} (1 - x_{new}^{(m)}) \, \underline{w}^{T}\underline{h}(Z^{(m)},\theta) \end{aligned}$$

$$\therefore \boxed{\frac{d\mathcal{L}}{d\gamma} = \sum_{m=1}^{M} (t^{(m)} - x_{new}^{(m)}(\gamma)) \cdot \underline{w}^{T}\underline{h}(Z^{(m)},\theta)}$$

(iii) The poor calibration of the original network is thought to be result of overfitting during the first training stage. Explain why the second stage of training is likely to improve the calibration of the network. [10%]

Answer
The first stage of learning can overfit as there are a large number of parameters in a CNN. This can lead to overconfident networks. The second stage is just fitting a single parameter to new data and therefore there is much less opportunity to overfit – it cannot explain every datapoint in the new dataset perfectly by only modifying a single parameter – so instead the training will refine probabilistic predictions so that they are well calibrated.

Engineering Part IIB 2018
Module 4F12 (Computer Vision) Assessor's Comments

1. **Gaussian smoothing and Harris corner detection**. Attempted by 71/78 Part IIB candidates, average mark 13.5/20.

   A very straightforward question covering convolution with low pass. The first part was well answered by most candidates. The second part on corner detection proved more challenging with many struggling to my the relationship between the auto-correlation matrix and the SSD of the two image patches.

2. **Perspective projection and camera calibration**. Attempted by 69/78 candidates, average mark 14.1/20.

   A question covering perspective projection. Well answered by most candidates. Marks were lost in the derivation of the 2 planes to define a ray - many missing the fact that the 2 planes are not-parallel and hence define a ray. Calibration with many noisy measurements were often missing details, i.e. the points spanning volume and not coplanar; least-squares formulation and RANSAC for outliers. The scaling due to depth was not well-explained in part d(i) although most were able to derive the weak perspective projection matrix.

3. **Projective transformations and epipolar geometry**. Attempted by 64/78 candidates, average mark 13.6/20.

   A well-answered question. Most marks were lost in part (b) on stated the epipolar matching constraint and deriving the equation of the line in view 2.

4. **Object rdetection with convolutional neural networks**. Attempted by 30/78 candidates, average mark 13.7/20.

   Many candidates that attempted this question made good progress.