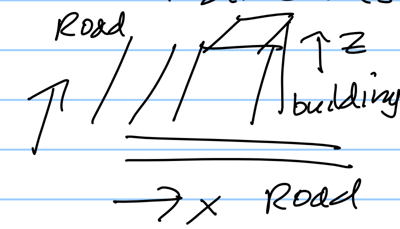
Lecture 5

Day 1

Manhattan World.

Many visual scenes have a Manhattan structure.

This give a natural 3-D coordinate system → y



A ground plane (x-y dimension) plus a vertical z-direction.

Most city (new cities) have this structure. Natural scenes – e.g. mountains and rivers – do not.

Two Questions:

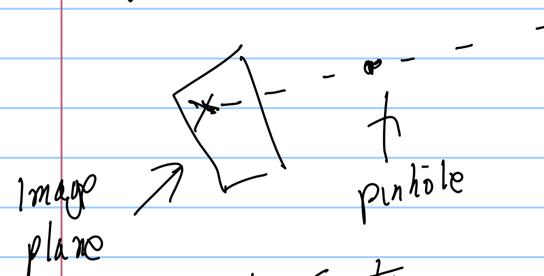
(i) How can a visual system use this knowledge to callibrate itself – ie. determine the angle of view compared to these x-y-z coordinates – and estimate these directions?

(ii) How can we decide if an image is Manhattan or not?

Note: the model we describe is not the best model to do this. But it is simple and gives important ideas.

Note: Humans see to assume a Manhattan structure. The Ames shows the visual illusions that happen if the image (ie. Ames room) appears to be Manhattan, but is not.

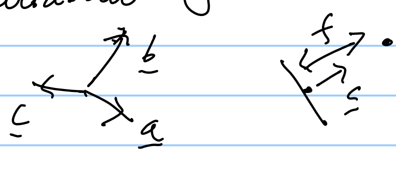
First, we need a model of projection. This is perspective projection.



x A point in space is projected by a light ray (straight line) which goes through the pin hole and is stopped when it hits the image plane.

(Note: real cameras have a small aperture and not a pinhole)

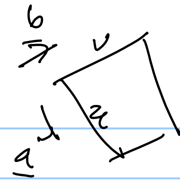
Camera Coordinate System



f focal length.

$\underline{c}$ direction of gaze.

$|\underline{a}| = |\underline{b}| = |\underline{c}| = 1$, $\underline{a}, \underline{b}, \underline{c}$ right angles page 2

Coordinates in image plane  Point $\underline{r}$ in space

Projection rules:

$$u = -f\frac{\underline{r}\cdot\underline{a}}{\underline{r}\cdot\underline{c}}, \quad v = -f\frac{\underline{r}\cdot\underline{b}}{\underline{r}\cdot\underline{c}}$$

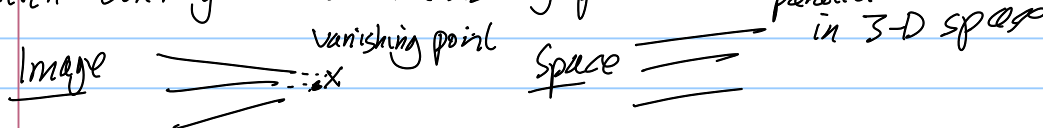
We specify $\Psi = (\underline{a}, \underline{b}, \underline{c})$

Camera has a 3-D coordinate system $\underline{a}, \underline{b}, \underline{c}$.

Manhattan world has a 3-D coordinate system x, y, z.

How to callibrate the camera by estimating the transformation between them?

The projection rules mean that parallel straight lines in the image will projection to straight lines in space which converge to a vanishing point



The position of the vanishing points in the image depend on the orientation Ψ of the camera with respect to the Manhattan coordinate system.

Parallel lines in x direction, vanishing points at $\left(-f\frac{a_x}{c_x}, -f\frac{b_x}{c_x}\right)$

" " " y " , " " " $\left(-f\frac{a_y}{c_y}, -f\frac{b_y}{c_y}\right)$

" " " z " , " " " $\left(-f\frac{a_z}{c_z}, -f\frac{b_z}{c_z}\right)$

where $\underline{a} = (a_x, a_y, a_z)$
$\underline{b} = (b_x, b_y, b_z)$
$\underline{c} = (c_x, c_y, c_z)$

Recall that $\underline{a}, \underline{b}, \underline{c}$ are orthogonal
$\underline{a}\cdot\underline{b} = \underline{a}\cdot\underline{c} = \underline{b}\cdot\underline{c} = 0$
$|\underline{a}| = |\underline{b}| = |\underline{c}| = 1$.

Now we introduce the model.

At each image pixel $\underline{u} = (u, v)$ there is a hidden variable $m_{\underline{u}}$ which indicates if the pixel is the image of an edge in the x, y, z directions, $m_{\underline{u}} \in \{1, 2, 3\}$,

an edge in a random direction, $m_{\underline{u}} = 4$

or not an edge $m_{\underline{u}} = 5$.

Let $\underline{E}_{\underline{u}}$ be the response of a derivative filter at $\underline{u}$

e.g. $\underline{E}_{\underline{u}} = \nabla I(\underline{u})$.

page 3

We put a prior probability on $m_{\underline{x}}$.

$$P(m_{\underline{x}}=1) = P(m_{\underline{x}}=2) = P(m_{\underline{x}}=3) = 0.02$$

$$P(m_{\underline{x}}=4) = 0.04, \quad P(m_{\underline{x}}=5) = 0.9$$

i.e. 90% of pixels in the image are not edges

In Manhattan 20% of edges are in the x, y, z directions (each); the remaining 40% of edges are randomly assigned.

$$P(\underline{E}_{\underline{x}} \mid m_{\underline{x}}) = P(|\underline{E}_{\underline{x}}| \mid m_{\underline{x}})\, P(\hat{\underline{E}}_{\underline{x}} \mid m_{\underline{x}})$$

$$\underline{E}_{\underline{x}} = |\underline{E}_{\underline{x}}|\, \hat{\underline{E}}_{\underline{x}}$$

magnitude; unit vector

For $m_{\underline{x}} = 1,2,3,4$

$P(|E_{\underline{x}}| \mid m_{\underline{x}})$ is specified by the models in lecture 3

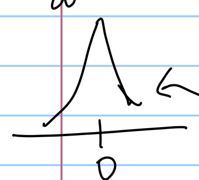
$P(|E_{\underline{x}}| \mid m_{\underline{x}}=5)$ by $P(|\underline{E}_{\underline{x}}| \mid \omega(x)=0)$ — no edge. i.e. $P(|E_{\underline{x}}| \mid \omega(x)=1)$ — edge

$P(\hat{\underline{E}}_{\underline{x}} \mid m_{\underline{x}})$ is uniform (all directions equally likely) for $m_{\underline{x}} = 4, 5$.

local image gradient direction

$$P(\hat{\underline{E}}_{\underline{x}} \mid m_{\underline{x}}) = F(|\hat{\underline{E}}_{\underline{x}} - \underline{n}(m_{\underline{x}}, \underline{\Psi}, \underline{x})|) \quad \text{for } m_{\underline{x}} = 1,2,3$$

where $\underline{n}(m_{\underline{x}})$ is the predicted direction of the edge — a function of $m_{\underline{x}}, \underline{\Psi}$ and $\underline{x}$.



$F(.)$ is peaked at 0, at predicted direction, but allows some variation.

Then
$$P(\underline{E}_{\underline{x}} \mid \underline{\Psi}, \underline{x}) = \sum_{m_{\underline{x}}=1}^{5} P(\underline{E}_{\underline{x}} \mid m_{\underline{x}}, \underline{\Psi}, \underline{x})\, P(m_{\underline{x}})$$

$m_{\underline{x}}$ are nuisance variables, so we sum them out. (standard Bayes)

For the entire image measurements
$$E = \{\underline{E}_{\underline{x}}\}$$

Assume
$$P(E \mid \underline{\Psi}) = \prod_{\underline{x}} P(\underline{E}_{\underline{x}} \mid \underline{\Psi}, \underline{x})$$
independence assumption

Note: this independence assumption is extremely unrealistic. Edges in the x-direction are usually continuous — so if $m_{\underline{x}} = 1$, then there is a higher probability that pixels near $\underline{x}$ also have $m_{\underline{x}} = 1$

local context

Really, we should have a model.
$$P(E \mid \underline{\Psi}) = \sum_{\{m_{\underline{x}}\}} \prod_{\underline{x}} P(\underline{E}_{\underline{x}} \mid m_{\underline{x}}, \underline{\Psi}, \underline{x})\, P(\{m_{\underline{x}}\})$$

where $P(\{m_{\underline{x}}\})$ is a prior which takes local context into account. But this model is harder to work with.

For the simple model (without the context)

$P(E|\Psi)$

we estimate $\hat{\Psi} = \underset{\Psi}{\mathrm{ARG\,MAX}}\ P(E|\Psi)$ (Can be done by exhaustive search)

This estimates the camera orientations relative to the Manhattan structure.

The results are good. We can display the estimated directions of the x, y, z directions in the images to compare with the true directions.

How to know if an image is Manhattan or not?

Answer is model selection –

We define an alternative null model for generating the image which does not assume Manhattan structure.

This null model is the same as the Manhattan model, but we set $P(m_u=1) = P(m_u=2) = P(m_u=3) = 0$.
$P(m_u=5) = 0.9$, $P(m_u=4) = 0.1$.

This removes any dependence on Ψ.

$$P_{null}(\underline{E}) = \prod_{u} \sum_{m_u} P(E_u|m_u)\, p(m_u)$$

Model selection:

Input image $I \to$ compute $\{E_u\}$

If $P_{null}(\{E_u\}) > P(\{E_u\}|\hat{\Psi}) + T$ ← best estimate of $\hat{\Psi}$ ← threshold

then image is not Manhattan.

Otherwise, image is Manhattan.

This gives good results → i.e.

classifies a city scene as Manhattan

classifies an image of fishes as non-Manhattan.