Vision Biologique et Artificielle :

Maths à Appliquer

Thierry Vieville

Résumé.

On cherche à modéliser comment un système biologique ou artificiel perçoit visuellement son mouvement propre et celui des objets de son environnement. On regarde aussi quelles sont les conséquences au niveau de la perception de la structure de cet environnement et comment il peut, en même temps, calibrer ses paramètres internes.

Ce sont des outils mathématiques qui nous permettent de modéliser ces procédés. Ils sont « presque » simples, et nous essayons de les explorer.

Introduction

Avant de présenter les outils mathématiques qui sont utilisés pour tenter de modéliser la perception visuelle, nous avons besoin de nous mettre d'accord sur quelques idées, mmm … disons épistémologiques(1).

Et pour cela, il nous faut une grenouille.

Et puis aussi une mouche.

Et ce qui est vraiment fascinant, voyez-vous, c'est que si la mouche vole à proximité de la grenouille … eh bien la grenouille gobera la mouche.

C'est même extraordinaire. Parce qu'une grenouille tout de même, c'est pas très malin. Quelques milliers de neurones tout au plus et pourtant…

Pourtant, le cerveau de la grenouille va faire des choses bien compliquées au demeurant :

– il va détecter parmi tous les objets de l'environnement ce qui « ressemble » à une mouche, disons, au moins à quelque chose qui bouge et qui passe à proximité,

– il va non seulement localiser la mouche là où elle est à l'instant où il la perçoit mais aussi prédire là où elle sera à la fin du saut de la grenouille … sinon ça loupera … puisque la mouche sera déjà passée !

Ainsi il a fallu modéliser non seulement quelques attributs géométriques de la mouche (sa position dans l'espace, sa taille peut-être) mais aussi la cinématique de sa trajectoire, et puis, et puis, il a fallu aussi modéliser la grenouille elle-même, la calibrer en fait : évaluer le délai qu'elle va mettre pour préparer un mouvement, estimer la « lenteur » du mouvement de la dite grenouille, pour qu'elle attrape finalement la mouche au bon endroit !

C'est pas si facile de gober une mouche, finalement.

Mais l'histoire n'est pas finie.

Allons chercher un caillou. Un caillou, la grenouille, elle s'en fout.

(1) Épistémologique : essayer d'apprécier la valeur de ces portions de sciences pour l'esprit humain (Larousse, 1970).

Mais si, d'aventure, on jetait ce caillou de manière telle que la trajectoire soit proche de celle de la mouche … alors … la grenouille goberait le caillou avec le même entrain.

Ce qui est un peu dommage pour la grenouille, mais c'est une excellente nouvelle pour nous.

Cela veut dire que, aussi complexe ce procédé soit-il de prime abord, il relève d'une « intelligence limitée », plus précisément, cela semble montrer que seuls les attributs géométriques et cinématiques sont pris en compte dans cette tâche perceptive, donc que leur modélisation mathématique est à notre portée.

C'est accessoirement aussi une très bonne nouvelle pour la mouche qui, on le constatera, a un peu tendance à « faire le caillou », bref à rester immobile pour échapper aux grenouilles. Ça nous permet en tout cas d'écraser plus facilement les mouches…

Eh bien, en vision biologique ou artificielle, nous modélisons précisément la perception des attributs géométriques et cinématiques des scènes visuelles observées. Cela ne nous permet pas de « comprendre » la scène. Mais tout de même d'y réaliser beaucoup de choses : détecter un objet en mouvement, évaluer sa distance relative, éviter des obstacles, poursuivre un objet mobile (genre une proie) ou s'alarmer d'être soi-même la cible (d'un prédateur), se localiser par rapport à des objets prédéfinis, combiner plusieurs vues d'un même objet en vue de calculer sa forme tri-dimensionnelle, etc.

Avec tout ça, il y a de quoi faire fonctionner un robot mobile, une grenouille, un bras articulé, un oiseau migrateur, une caméra qui détecterait un début d'incendie, le galop d'un cheval, bref effectuer des tâches visuo-motrice biologiques ou artificielles.

On est, certes, loin de l'analyse « sémantique » d'une scène par la vision humaine ! Mais trouver du « sens » à quelque chose est un autre problème…

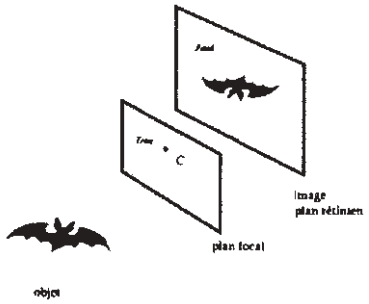
Essayons avec les grenouilles… C'est déjà passionnant.

Des caméras aux boites de carton

Pour ceux qui essayent de trouver des équations permettant de modéliser la vision, les caméras ou les yeux se comportent comme … une boite en carton.

Avec un trou pour laisser passer la lumière.

Avec un fond pour recevoir l'image.



Pour un plan image situé à une distance « focale » f le long de l'axe des **Z**,
un point de l'espace 3D $\mathbf{M} = (X, Y, Z)$ se projette sur un point de l'image 2D $\boldsymbol{m} = (u, v, f)$ défini par :

$$\begin{cases} u = f\ X / Z \\ v = f\ Y / Z \end{cases}$$

car $\boldsymbol{m}$, **M** et le centre optique origine **C** sont sur un même rayon optique rectiligne.

La lumière passe par le « trou » et vient se projeter sur le « fond » : on parle d'un sténopé.

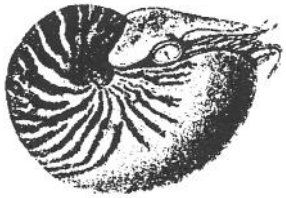
D'après des écrits retrouvés en Chine et datant du Ve siècle avant Jésus-Christ, on pense que le phénomène de formation des images sténopées a été observé dès cette époque.

À la Renaissance on retrouve le début d'une utilisation de caméras sténopées aussi bien au niveau artistique que scientifique (en astronomie, pour l'observation des éclipses).

Ces caméras sont encore utilisées au XXe siècle pour des applications bien spécifiques en physique nucléaire telles que l'observation de rayons en provenance du soleil ou des rayons de hautes énergies dans les plasmas laser.

Schéma de caméra sténopée, De Radio Astronomica et Geometrica, par Gemma Frisius, 1545, utilisée pour l'observation de l'éclipse de soleil de 1544.

Il y a même une espèce biologique utilisant ce principe, c'est le mollusque « Nautilus » dont l'œil est un trou à ouverture réglable.



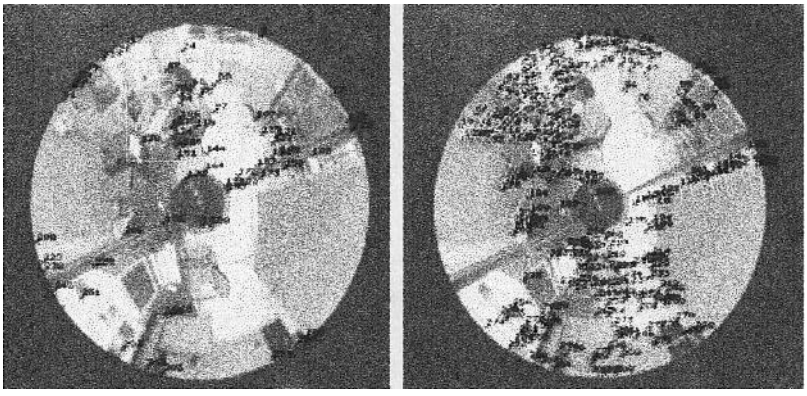
L'oeil du Nautilus fonctionne selon le principe sténopé

Les mathématiciens, eux, y voient autre chose.

Ils ne regardent pas le fond de la boite mais … les rayons lumineux qui passent par le trou.

Car ce qui importe ici, ce sont les *rayons lumineux*, chacun n'étant défini que par son orientation gauche-droite et haut-bas, bref deux paramètres : le « faisceau » de droites qui passent par le centre optique.

Parce que le fond de la boite pourrait être planaire, sphérique ou de toute autre forme régulière, il y aurait toujours une « image » à regarder, quoique « tordue ».



Une caméra omni-directionnelle produit une image tordue, mais permet de regarder … partout !

Considérer l'intersection d'une partie de ces droites avec une surface « rétienne » et regarder le point planaire qui y correspond n'est qu'un moyen de choisir ces deux paramètres.

On dit que cette « image » constitue le plan projectif $\mathcal{P}^2$.

Disons que le monde est rigide

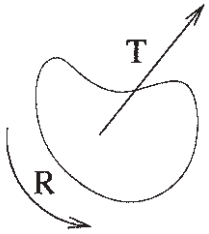
Pour « voir » ce qui bouge dans le monde, il faut faire quelques hypothèses.

D'abord on suppose que *le monde est essentiellement fait d'objets rigides*, c'est à dire qui ne se déforment pas.

• C'est le cas du sol, des bâtiments, d'un véhicule et de tous les objets immobiles.

• C'est aussi *presque* le cas d'un corps humain : il est « rigide par morceaux », autrement dit fait d'objets rigidement reliés entre eux.

• Ce n'est pas le cas d'un arbre sous le vent,d'une rivière, sauf si on regarde d'assez loin et *néglige les déformations*, ce que nous ferons ici.



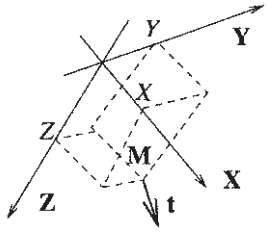
La localisation d'un objet dans l'espace est définie par
(a) trois valeurs de position
et
(b) trois angles pour son orientation.

Une scène est vue comme un ensemble d'objets rigides.

Pour un objet rigide, son mouvement se décompose en :

• (1) une *translation*, quand sa *position* bouge, notée $\boldsymbol{t}$ et définie dans les trois directions de l'espace,

• (2) une *rotation* (de l'objet autour de lui même), quand son *orientation* bouge, notée $\boldsymbol{r}$, à trois paramètres.



Comment *calculer* un mouvement rigide ?

Prenons un point sur cet objet **M** = (X, Y, Z) repéré par ses positions le long des axes **X**, **Y** et **Z**.

On voit bien qu'au cours d'une translation $\boldsymbol{t} = (t_x, t_y, t_z)$, M bouge justement de $\dot{\mathbf{M}} = \boldsymbol{t}$.

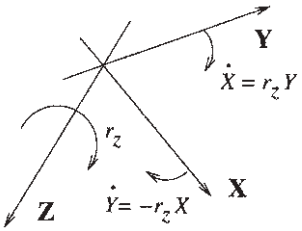
Mais que se passe-t'il pour une rotation ? C'est un peu plus compliqué…

Pour une rotation r_z autour de l'axe des **Z**, regardons à droite, ici :

• il y a une « action croisée », un « bras de levier ».

• un point sur l'axe **Y** va se décaler le long de l'axe des **X**, ceci proportionnellement à r_z ET l'éloignement Y, comme représenté ici.

• un point sur l'axe **X** va se déplacer le long de l'axe des **Y**, au signe près, de la même façon.



En rassemblant ces calculs pour les trois axes, on obtient :

$$\begin{cases} \dot{\mathrm{X}} = t_x + r_z \mathrm{Y} - r_y \mathrm{Z} \\ \dot{\mathrm{Y}} = t_y + r_x \mathrm{Z} - r_x \mathrm{Y} \\ \dot{\mathrm{Z}} = t_z + r_y \mathrm{X} - r_x \mathrm{Y} \end{cases}$$

que l'on note aussi :

$$\dot{\mathbf{M}} = \boldsymbol{t} + \boldsymbol{r} \wedge \mathbf{M}.$$

On parle ici de *torseur de vitesse*.

L'équation fondamentale de la perception du mouvement

Considérons un point 3D qui a un mouvement rigide et sa projection 2D dans l'image.

En dérivant:

$$\begin{cases} u = f\ \mathrm{X}/\mathrm{Z} \\ v = f\ \mathrm{Y}/\mathrm{Z} \end{cases}$$

avec la relation précédente, on réalise un long et fastidieux … mais très simple exercice de calcul qui donne ceci :

$$\begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix} = \frac{1}{\mathrm{Z}} \begin{bmatrix} f & 0 & -u \\ 0 & f & -v \end{bmatrix} \begin{bmatrix} t_x \\ t_y \\ t_z - \mathrm{Z}\dfrac{\dot{f}}{f} \end{bmatrix} + \frac{1}{f} \begin{bmatrix} -uv & u^2 + f^2 & -fv \\ -v^2 - f^2 & uv & -fu \end{bmatrix} \begin{bmatrix} r_x \\ r_y \\ r_z \end{bmatrix}$$

$$\dot{\boldsymbol{m}} = \Pi \qquad \mathbf{A} \qquad \boldsymbol{t}' \qquad + \qquad \mathbf{B} \qquad \boldsymbol{r}$$

Ici $\dot{\boldsymbol{m}}$ est le mouvement de la projection 2D du point 3D dans l'image. Il dépend :

* de la « proximité » ÷ du point 3D par rapport à l'image, l'inverse de sa « profondeur » ;
* du mouvement 3D du point : rotation $\boldsymbol{r}$ et translation $\boldsymbol{t}'$ (avec un terme en plus qui dépend du zoom) ;
* de la position 2D (u, v) dans l'image et de la focale f présentes dans **A** et **B**.

Et l'analyse de cette équation est incroyablement riche en enseignement.

Un voyage en translation

Disons qu'il n'y a qu'une translation horizontale et pas de rotation ...

Parallaxe : s'il n'y a qu'une translation horizontale t_x, le mouvement est de la forme :

$$\dot{u} = f\,\frac{1}{\mathrm{Z}}\,t_x$$

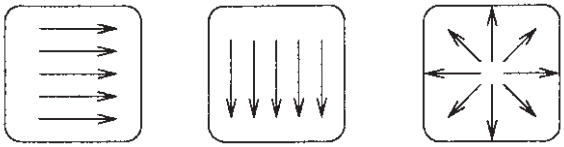
C'est à dire que le mouvement a la même forme partout sur la rétine !

→ On peut alors évaluer la proximité absolue $\Pi = \frac{1}{Z}$ si on connaît la translation t_x et la focale f ;
→ On peut aussi évaluer la proximité *relative* entre deux points d'un même objet rigide puisque $\frac{\dot{u}}{\dot{u}'} = \frac{\Pi}{\Pi'}$.

De Liliput à Brobdingag : avec un œil on ne voit PAS la taille du monde !
→ Si nous multiplions la translation $\boldsymbol{t}$ et toutes les profondeurs Z par le même facteur, l'équation reste *invariante* ! Elle ne « bouge » pas, donc impossible de détecter ce « facteur d'échelle », sauf avec deux yeux.

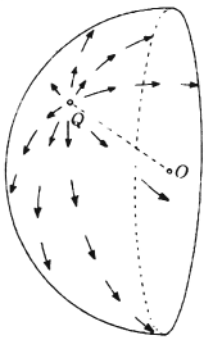
Où est l'horizon ? Dans la même équation ! Une image est faite de points et on ne peut détecter de mouvement inférieur à une valeur unité, donc si $\dot{u} < 1 \Rightarrow Z > Z_\infty = f\, t_x$, il n'y a plus de mouvement !!!
• (1) L'horizon est, pour une caméra, un plan d'équation $Z > Z_\infty$.
• (2) L'horizon « recule » si (i) la focale augmente (si on « zoome ») ou si (ii) la translation augmente.
• (3) S'il n'y a pas de translation, on ne perçoit aucune profondeur Z : tout est à l'horizon !

Regardons ce qui se passe avec des translations plus compliquées.



t_x / Z t_y / Z $t_z / Z - \dot{f} / f$

Bien sûr avec une translation verticale on observe des phénomènes similaires.



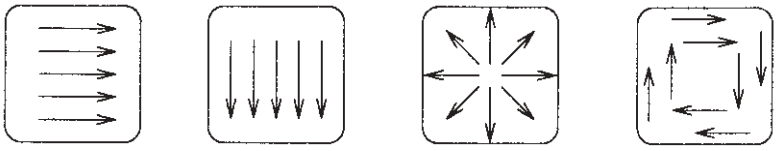
Coup de boule : s'il y a une translation t_z en profondeur :

$$\begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix} = \left[\frac{\dot{f}}{f} - \frac{t_z}{Z} \right] \begin{bmatrix} u \\ v \end{bmatrix}$$

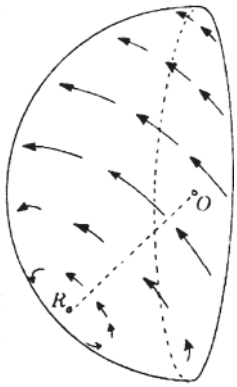
→ au centre $(u, v) = (0, 0)$ de la rétine, on ne voit plus rien !!!
← Mais ce point « singulier » donne directement la direction de la translation.
→ On ne peut pas non plus distinguer une variation de focale (« zoom ») de la translation d'un plan fronto-parallèle (de profondeur constante).

Quand la rotation s'en mêle ..

Si la rotation arrive, tout se complique … et s'améliore !



$t_x / Z + r_y$ $\quad$ t_y / Z $\quad$ $t_z / Z - \dot{f} / f$ $\quad$ r_z

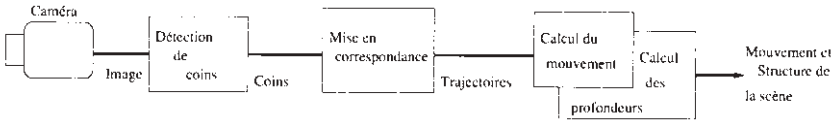


→ Au centre de la rétine les rotations r_x et r_y « ressemblent » à des translations, mais ce n'est pas le cas de la rotation r_z qui peut être utilisée pour mesurer, même lors de translations, une orientation comme un « volant ».

→ La rotation ne dépend PAS de la profondeur : c'est une simple transformation de l'image !

⇒ On peut donc décaler l'image en rotation et « recaler » les deux images, sans parallaxe, quelle que soit la scène : de l'avantage d'avoir les yeux « ronds ». Coïncidence ?

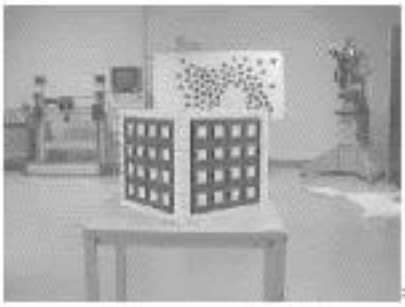
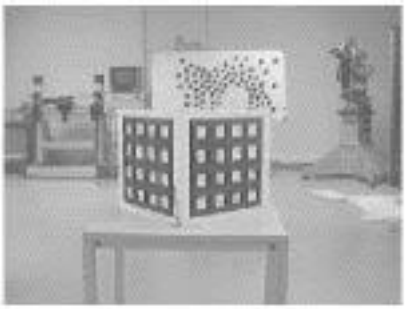
Des équations à un module informatique



Comment passer de ces équations à un programme qui analyse une vidéo ?

→ On détecte des points d'intérêts (des coins) dans chaque image, en regardant les zones de forte courbure.

→ On met en correspondance ces points dans une séquence d'image en comparant autour de chaque point une petite zone de l'image, choisissant les points dont les zones se ressemblent le plus.

→ On calcule alors le mouvement dans l'image et grâce à ces équations en :
(1) détectant les différents mobiles de la scène (ceux qui ont le même mouvement),
(2) calculant les tailles et proximités relatives des objets (ou ceux qui sont à l'horizon).
Pour suivre, détecter, attraper un objet … comme une grenouille bien intentionnée !