Perception du mouvement: notions fondamentales

Thierry.Vieville@inria.fr

janvier 2001

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 .. et 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 !

1. Epistémologique : essayer d'apprécier la valeur de portions de sciences pour l'esprit humain *Larousse, 1970*

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 cailloux. Un cailloux, la grenouille, elle s'en fout.

Mais si d'aventure, on jetait ce cailloux de manière telle que la trajectoire soit proche de celle de la mouche .. alors .. la grenouille goberait le cailloux 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 cailloux", bref rester immobile pour échapper aux grenouilles. Ça nous permet en tout cas d'écraser plus facilement les mouches ..

Et 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 tridimensionnelle, 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 ..

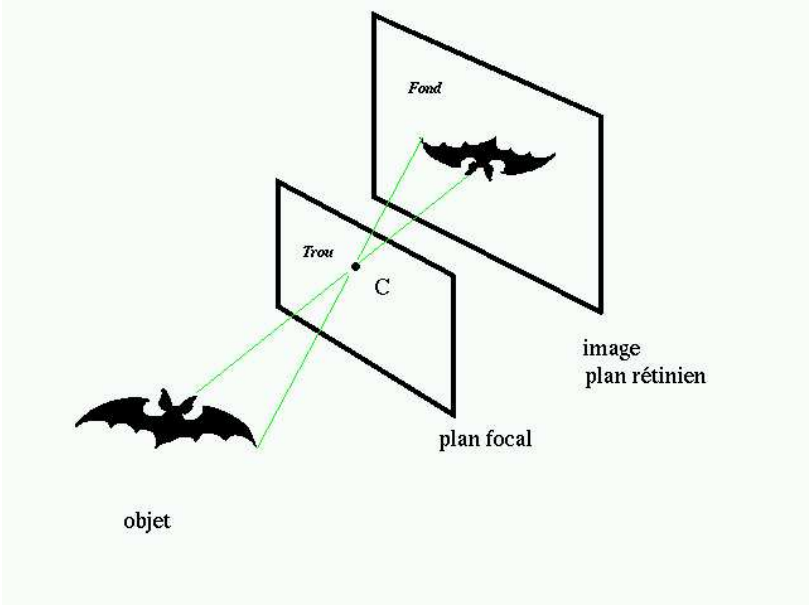
Essayons avec les grenouilles .. c'est déjà passionnant.

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 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 $\mathbf{Z}$, un point de l'espace 3D $\mathbf{M} = (X,Y,Z)$ se projette sur un point de l'image 2D $\mathbf{m} = (u,v,f)$ défini par:

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

car $\mathbf{m}$, $\mathbf{M}$ et le centre optique origine $\mathbf{C}$ 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 Vème siècle avant Jésus-Christ, on pense que le phénomène de formation des images sténopées a été observé dès cette époque.

A 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).

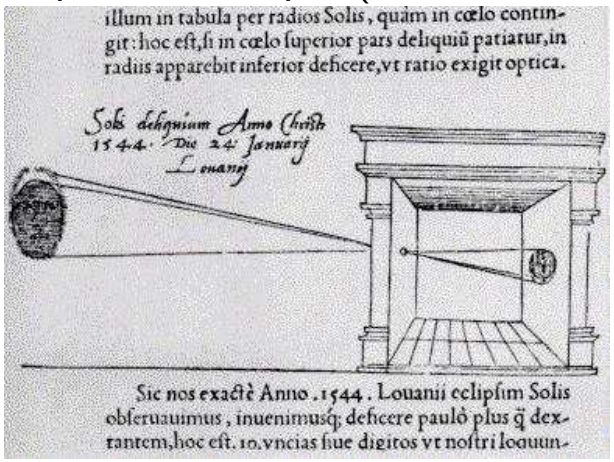
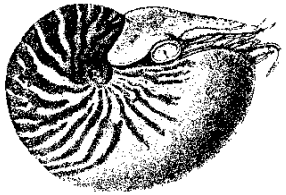


Schéma de caméra sténopée,
De Radio Astronomica et Geometrica, par Gemma Frisius, 1545,
utilisée pour l’observation de l’éclipse de soleil de 1544.

Ces caméras sont encore utilisées au XXème 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.

Il y a même une espèce biologique utilisant ce principe, c’est le mollusque “Nautilus” dont l’oeil est un trou à ouverture réglable:

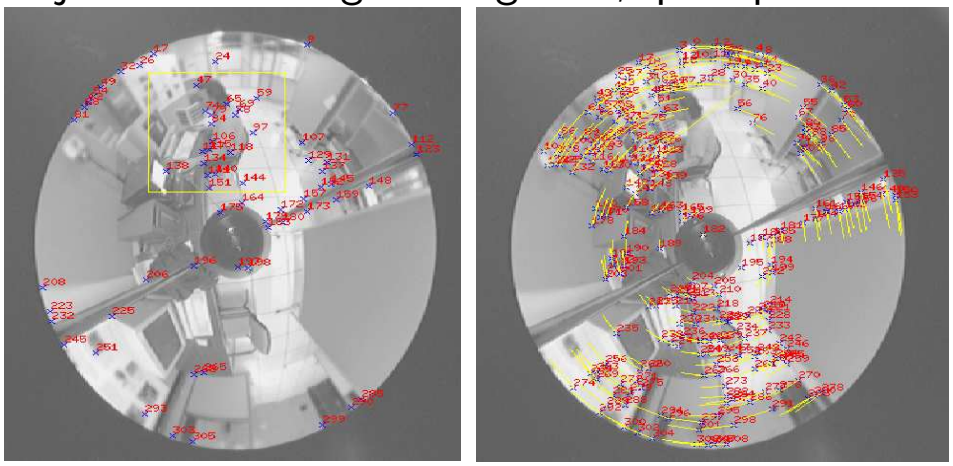


L’oeil du Nautilus fonctionne
selon le principe sténopé.

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 2 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, quoi que "tordue".



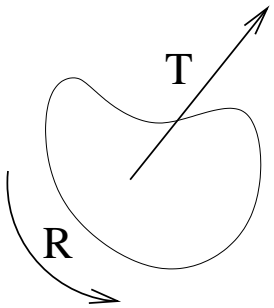
Une caméra omnidirectionnelle
produit une image tordue
mais permet de regarder .. partout !

Considérer l'intersection d'une partie de ces droites avec une surface "rétinienne" 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
. .. *négligent les déformations* .. .. ce que nous ferons ici.

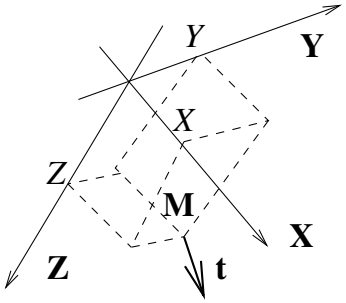


La localisation d'un objet dans l'espace est définie par
(a) 3 valeurs de position et
(b) 3 angles pour son orientation.

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 $\mathbf{t}$ et définie dans les 3 directions de l'espace,
. (2) une *rotation* (de l'objet autour de lui même), quand son *orientation* bouge, notée $\mathbf{r}$, à 3 paramètres.

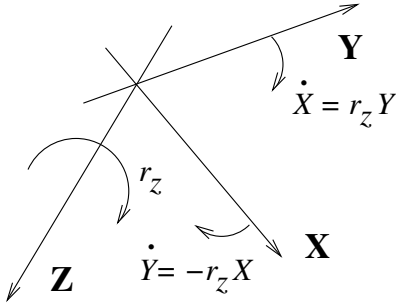
Comment *calculer* un mouvement rigide ?



Prenons un point sur cet objet $\mathbf{M} = (X,Y,Z)$
repéré par ses positions le long des axes $\mathbf{X}$, $\mathbf{Y}$ et $\mathbf{Z}$.
On voit bien qu'au cours d'une translation $\mathbf{t} = (t_x,t_y,t_z)$,
$\mathbf{M}$ bouge justement de $\dot{\mathbf{M}} = \mathbf{t}$.

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 $\mathbf{Z}$, regardons à droite, ici -¿
.. il y a une "action croisée" .. un "bras de levier" ..
- un point sur l'axe $\mathbf{Y}$ 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 $\mathbf{X}$ va se déplacer le long de l'axe des Y,
au signe près, de la même façon.



En rassemblant ces calculs pour les 3 axes, on obtient:

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

que l'on note aussi $\dot{\mathbf{M}} = \mathbf{t} + \mathbf{r} \wedge \mathbf{M}$

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: $\left\{ \begin{array}{rcl} u & = & f\,X/Z \\ v & = & f\,Y/Z \end{array} \right.$ avec la relation précédente ..

.. on réalise un long et fastidieux .. mais très simple exercice de calcul qui donne ceci :

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

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

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

. * de la "proximité" Π du point 3D par rapport à l'image, l'inverse de sa "profondeur"

. * du mouvement 3D du point : rotation $\mathbf{r}$ et translation $\mathbf{t}'$ (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 $\mathbf{A}$ et $\mathbf{B}$.

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:

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

+ C'est à dire que le mouvement a la même forme partout sur la rétine !
$\longrightarrow$ on peut alors évaluer la proximité absolue $\Pi = \frac{1}{Z}$ si on connaît la translation t_x et la focale f
$\longrightarrow$ on peut aussi évaluer la proximité *relative* entre deux points d'un même objet rigide puisque $\dot{u}/\dot{u}' = \Pi/\Pi'$

De Liliput à Brobdingag: avec un oeil on ne voit PAS la taille du monde !
$\longrightarrow$ si nous multiplions la translation $\mathbf{t}$ 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 2 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

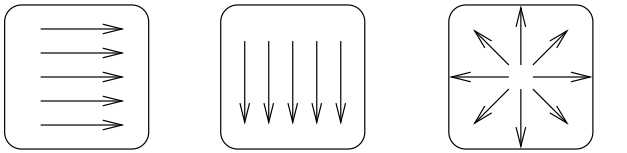
de mouvement !!!

. (1) l'horizon est, pour une caméra, un plan d'équation $Z > Z_\infty$

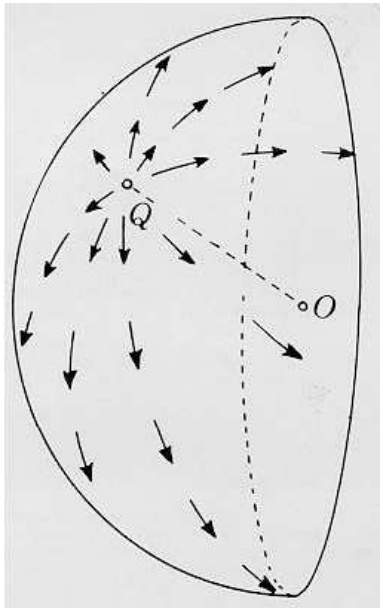
. (2) l'horizon "recule" si (i) la focale augmente (si on "zoom") ou si (ii) la translation augmente ..

. (3) si 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 ..



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



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

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

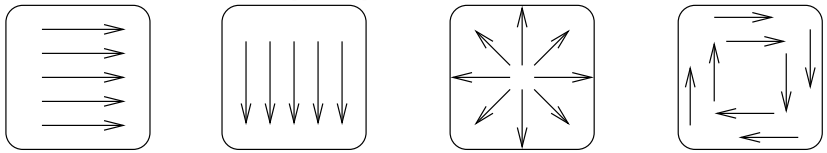
$\longrightarrow$ au centre $(u,v) = (0,0)$ de la rétine .. on ne voit plus rien !!!

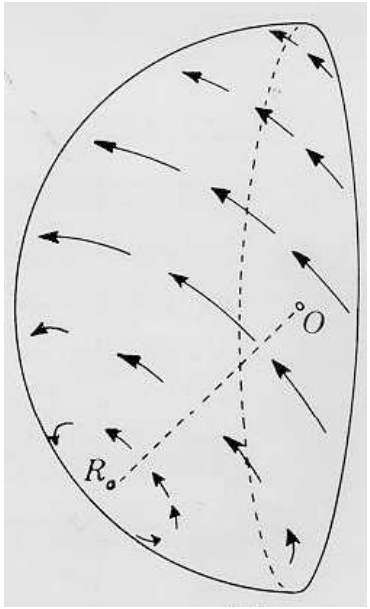
$\longleftarrow$ Mais ce point "singulier" donne directement la direction de la translation.

$\longrightarrow$ 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 !





$\longrightarrow$ Au centre de la rétine les rotations r_x et r_y "ressemblent" à des translations ..

.. mais ce n'est pas le cas de rotation r_z qui peut être utilisée pour mesurer, même lors de translations, une orientation comme un "volant"

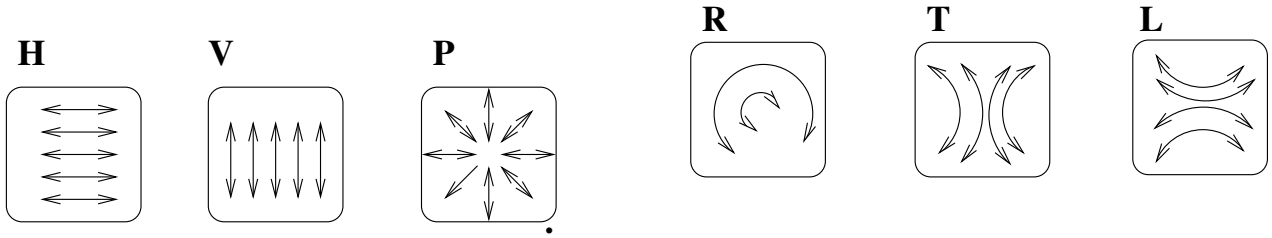
$\longrightarrow$ La rotation ne dépend PAS de la profondeur ..

.. c'est une simple transformation de l'image !

$\Rightarrow$ on peut donc décaler l'image en rotation et ..

.. "recaler" les deux images, sans parallaxe, quelque soit la scène .. .. de l'avantage d'avoir les yeux "ronds" .. coïncidence ?

Les 3 types de mouvements



Mouvements panoramiques. Les composantes panoramiques permettent de mesurer la profondeur relative des points d'une scène en utilisant le fait que *les objets proches bougent en valeur relative plus que les objets lointains* (c'est l'effet de parallaxe). Ces mouvements panoramiques permettent aussi de suivre visuellement un objet en mouvement en effectuant une poursuite oculaire.

Mouvement en profondeur. Les composantes en profondeur permettent de détecter une cible qui s'éloigne (par exemple une "proie") ou qui se rapproche (par exemple un "prédateur") et de mesurer le "temps de collision", c'est à dire le délai nécessaire pour que la distance en profondeur de la cible à la caméra s'annule.

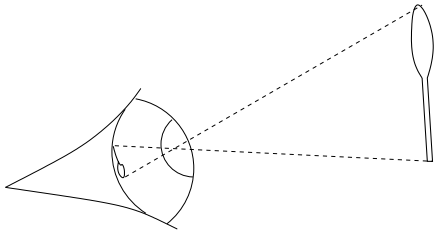
Mouvements angulaires. Les composants angulaires peuvent être facilement compensées, ce qui permet de stabiliser l'image d'une vue dynamique. En vision biologique, un de ces mécanismes de compensation est le réflexe opto-cinétique. Il est souvent associé aux capteurs de force, dits inertiels, qui se trouvent dans l'oreille interne des animaux et des humains et qui forment le système vestibulaire. Le réflexe vestibulo-oculaire qui lui est associé permet de mesurer les accélérations angulaires et linéaires de la tête pour stabiliser le regard. Les termes hyperboliques liés à ces mouvements permettent la *segmentation* des différentes zones de mouvement

Un capteur visuel se calibre lui-même.

Lors des mouvements angulaires de l'oeil, la vue de la scène n'est pas déformée sur la rétine: on regarde simplement la scène sous un autre angle, simple "déplacement angulaire" de l'image.

Si au cours d'un déplacement angulaire il apparaît une déformation inattendue de l'image, c'est que les paramètres géométriques ou optiques du capteurs sont erronés.

En les modifiant de façon à réduire la déformation observée, il est possible de calibrer la caméra.



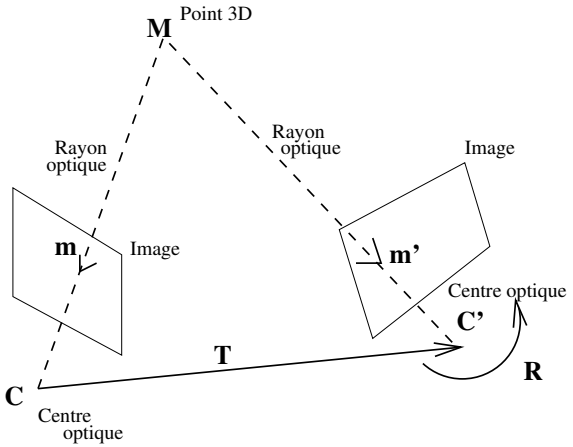
$\Rightarrow$ Utilisable avec d'autres déplacements que des rotations.

Serait-ce la raison pour laquelle l'oeil est-il rond ?

L'oeil est en effet le seul organe moteur qui ait une telle régularité géométrique chez presque tous les animaux. La nécessité de réaliser des rotations exactes de l'oeil lors du passage d'une vue à l'autre viendrait de la nécessité de pouvoir recoller les deux vues (sans tenir compte de la profondeur des objets, leur éventuelle occlusion, etc..) par un simple "glissement angulaire" d'image.

En faisant des rotations pures, l'oeil se construit une vue panoramique de son environnement. On utilise ce même procédé en vision artificielle pour construire de grandes vues composites (on parle de mosaïques) à partir de plusieurs photos.

Equivalence avec la stéréoscopie



Percevoir la profondeur grâce à la géométrie.

En stéréoscopie, on considère deux caméras dont les centres optiques sont notés **C** *et* **C'** *et dont la position relative est spécifiée par une translation* **T** *et l'orientation relative par une rotation* **R**.

Pour un point de l'espace **M**, *le fait que le rayon optique passant par* **m** *issu de* **C** *et le rayon optique passant par* **m'** *issu de* **C'** *se coupent en* **M** *induit une contrainte géométrique, dite "épipolaire".*

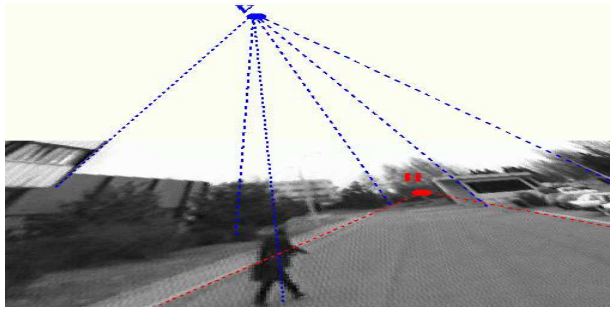
En considérant les équations correspondant à ces contraintes épipolaires pour tous les points de la scène observée (i) la distance absolue entre les points observés et les caméras peuvent être calculées, tandis que (ii) les positions et orientations relatives des caméras et les paramètres optiques de deux caméras peuvent être "auto"-calibrées en partie.

Vision et orientation spatiale

Le monde visuel est globalement *rigide et stationnaire* (mouvement propre) les objets mobiles sont détectés comme artefacts.

Les points infiniment lointains, à l'horizon, restent *immobiles en translation* (points de fuite, directions de l'espace).

La direction verticale permet la gyro-rotation ("roulis" + "tangage").



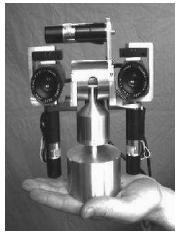
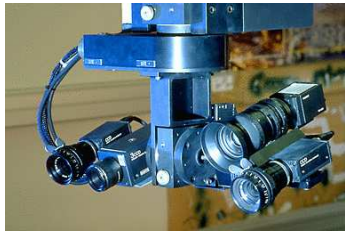
Détection de l'horizon et de la verticale : les directions moyennes verticales des bords de bâtiments, arbres, etc. d'une part et des bords de la route, supposée droite, d'autre part sont des "parallèles" qui se coupent à "l'infini", c'est à dire l'horizon.

$\Rightarrow$ **la vision fournit des informations inertielles**

Vision et contrôle du regard

Les objets d'intérêts sont les objets *proches*, *mobiles ou texturés*.

Le contrôle oculomoteur (*saccades/poursuite en interaction*, *stabilisation visuo/inertielle*) est utilisé tel-quel en robotique.

L'analogie avec les systèmes biologiques a conduit à la conception de caméras dotées des mêmes fonctionnalités que les yeux. Leurs déplacements permettent de contrôler la direction du regard pour explorer l'environnement visuel, stabiliser la vision si le sujet se déplace, suivre un objet qui bouge, etc..

Bien entendu, ces systèmes robotiques ne sont pas anthropomorphiques.

En vidéo-conférence ou en télésurveillance les mécanismes sous-jacents sont désormais couramment utilisés.

Vision et modèles de l'environnement

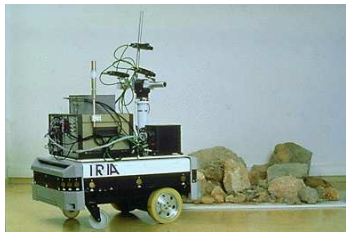
Les modèles sont dédiés à une unique tâche visuo-motrice.

Ces modèles paramétriques détectent des artefacts.

Ces modèles testent leur validité (par rapport à un autre modèle).

Ex: mur gauche + mur droit + sol $\Rightarrow$ navigation 3d autonome

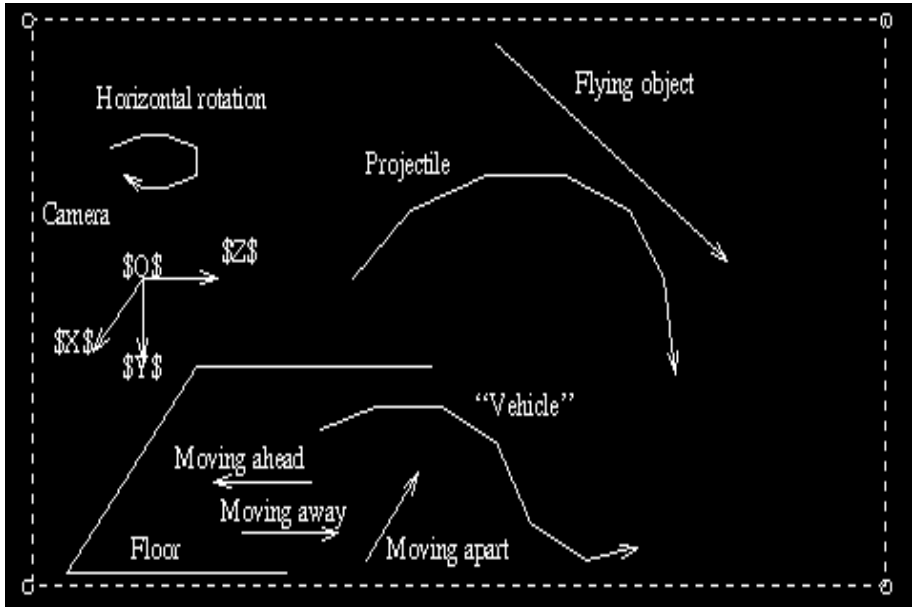
Ex: élément supplémentaire au modèle $\Rightarrow$ obstacle !

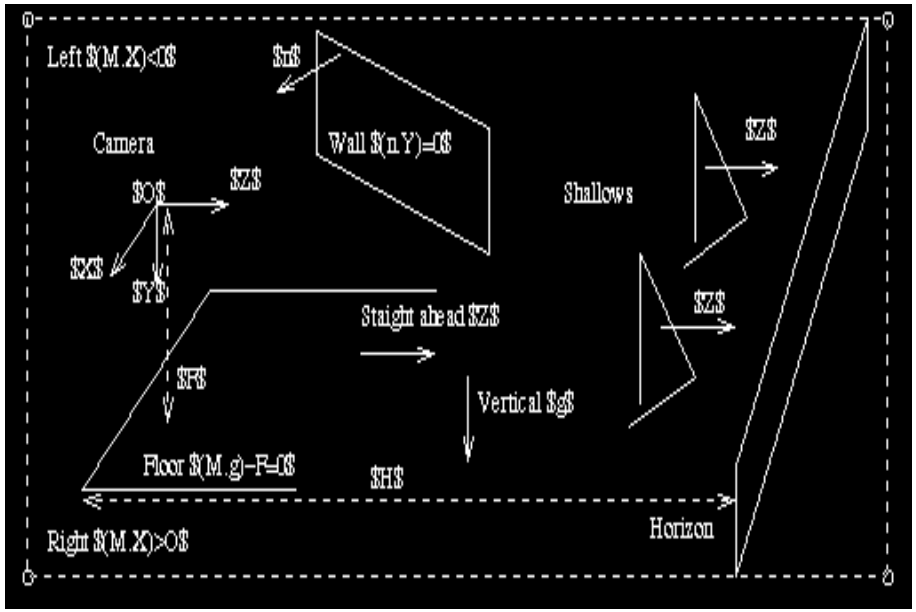



Un système visuel 3D rudimentaire à trois caméras permet à un robot mobile d'éviter une personne. Il ne différencie évidemment pas les obstacles entre eux (humain, mur, ..) mais se contente de détecter les bords gauche et droit de son "couloir" de navigation.

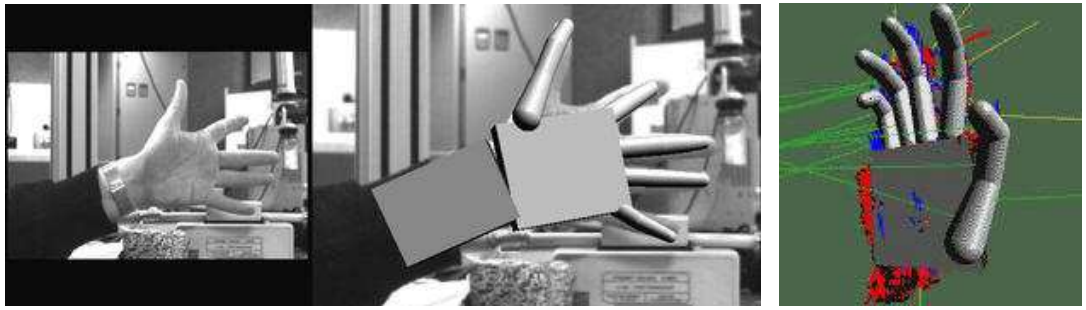
Ces modèles s'insèrent dans une hiérarchie (rasoir d"Occam).

Le test d'hypothèses permet de construire une certaine "sémantique" dans des univers limités.



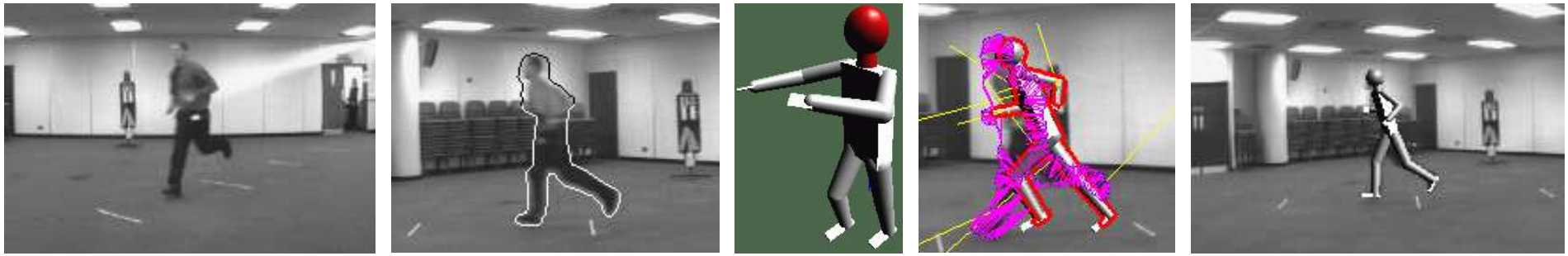


.. jusqu'à des mécanismes plus évolués de perception.



L'utilisation de modèles a priori permet de détecter et d'observer des objets complexes (ici les coordonnées articulaires d'un avatar) en les représentant par un faible nombre de paramètres pertinents.

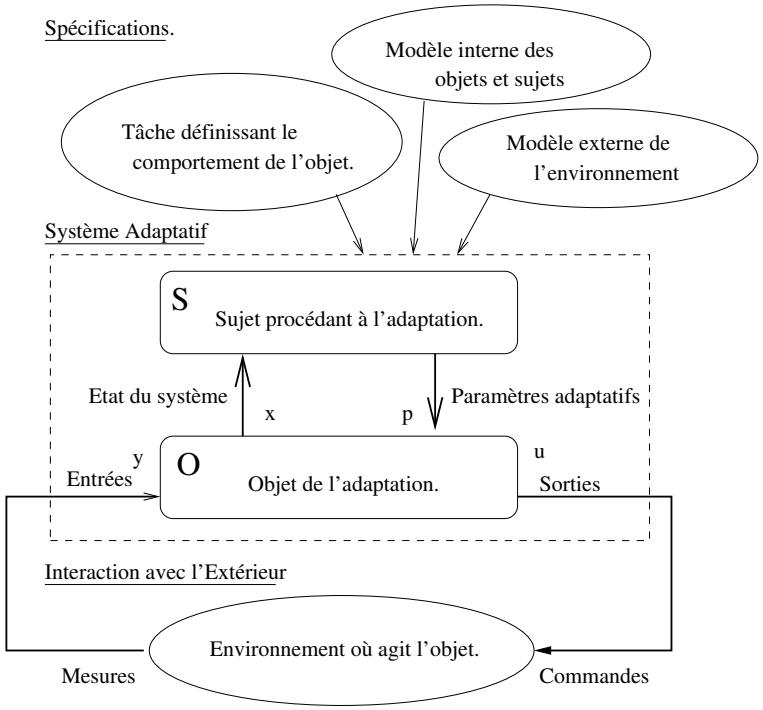
C'est l'adaptation de ces paramètres qui permet d'estimer les propriétés nécessaires à la reconnaissance de cette main et à son suivi au fur et à mesure du déroulement d'un geste.



Fédérer ces concepts sous la notion d'adaptation

L'adaptation comme un choix d'architecture.

L'adaptation comme un mécanisme d'apprentissage paramétrique.



Les cinq clés de l'adaptation. Pour le dictionnaire, adapter veut dire changer un "comportement" pour lui permettre de réagir dans de nouvelles "circonstances". Plus précisément il y a deux acteurs: (1) l'objet, c'est à dire ce qui est adapté (c'est souvent un mécanisme, un instrument, un dispositif, un réflexe ou un comportement) et (2) le sujet, c'est à dire le mécanisme qui adapte.

Mais une telle définition n'est pas suffisante. Prenons, par exemple, le cas d'un animal qui change de conditions de vie. Il est clair qu'il doit "adapter" son comportement au nouvel environnement dans lequel il est plongé. C'est ce qu'il fera, mais à partir de ce qu'il savait déjà faire avant c'est à dire d'un modèle de référence. Bien sûr, ceci a lieu seulement si il y est obligé par une variation de son environnement. D'autre part, il le fera dans un but précis, sa survie en l'occurrence.

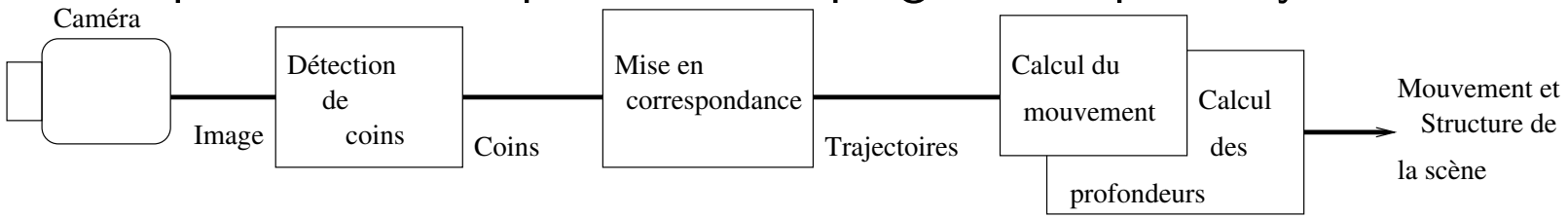
Dans un processus d'adaptation, il y a donc en plus d'un objet et d'un sujet: un modèle de référence qui permet de passer d'un fonctionnement habituel et à un nouveau fonctionnement et une tâche à accomplir qui motive le passage d'un comportement à un autre.

Tout ceci a lieu dans un environnement dont une modification suscite le processus d'adaptation.

Il y a finalement cinq éléments en jeu et il semble raisonnable d'énoncer: l'adaptation correspond à un processus par lequel un *sujet*, lorsqu'il enregistre une variation de l'*environnement*, modifie les paramètres d'un *objet*, à partir d'un *modèle de référence*, dans le but d'accomplir une *tâche perceptive spécifique*.

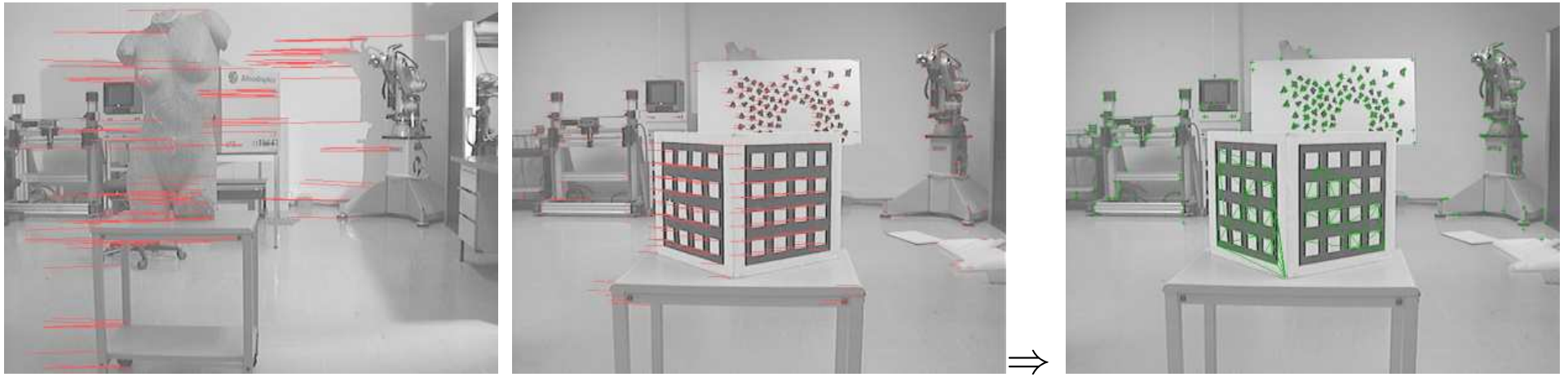
Des équations à un module informatique ..

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



$\longrightarrow$ On détecte des points d'intérêts (des coins) dans chaque image,
$- - - >$ en regardant les zones de forte courbure.

$\longrightarrow$ 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.



$\longrightarrow$ On calcule alors le mouvement dans l'image et grâce à ces équations :

(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 !

Conclusion

En cherchant à modéliser comment un système biologique ou artificiel perçoit son mouvement propre et celui des objets de son environnement et g quelles en sont les conséquences au niveau de la perception de la structure de cet environnement et de l'auto-calibration de ses paramètres internes.

Cet axe de recherche a permis, en vision artificielle :

- d'introduire l'usage de mesures inertielles, en particulier l'auto-calibration d'un tel capteur, en fusion avec les informations visuelles,
- de proposer des méthodes d'auto-calibration actives (et en un sens optimales) des paramètres du capteur visuels, y compris ceux des mécanismes de vision précoce,
- de démontrer qu'une analyse hiérarchique des différents modèles spécifiques de mouvement pouvant correspondre à des contraintes mécaniques ou approximer une situation de mouvement réel permet de procéder à une paramétrisation optimale du phénomène observé,

tandis que ce travail est aujourd'hui complété par (i) une étude sur le contrôle de la focale d'une caméra et les mécanismes de focalisation d'attention liés à ce processus, (ii) une expérimentation sur la validité de ces modèles pour rendre compte de la perception biologique du mouvement.