Estimation de distance et calibrage de caméra

Les objectifs du projet sont à termes de pouvoir estimer la profondeur d'un objet sur une image et de pouvoir calibrer la caméra utilisée dans le cadre d'images professionnelles et non professionnelles.

En calibrant une caméra on peut par la suite déterminer de nombreux paramètres autres que la position, telle que la vitesse d'un objet sur une vidéo ou sa taille.

L'étalonnage de caméra est la porte d'entrée à une multitude d'application aussi bien dans l'optique, le cinéma ou encore la vision par ordinateur.

Le calibrage géométrique (ou étalonnage géométrique) d'une caméra permet d'estimer les paramètres de l'objectif et du capteur d'image d'une caméra d'images ou vidéo. Cela consiste à déterminer la relation mathématique entre les coordonnées des points 3D de la scène observée et les coordonnées 2D de l'image.

Lors du passage d'une scène 3D à une scène 2D de nombreuses informations sont perdues telles que la profondeur ou la perspective. Un phénomène de distorsion accompagne aussi le passage de la scène 3D à l'image. Le calibrage de caméra consiste donc à retrouver cette transformation mathématique qui s'opère lors du passage de la scène à une image 2D.

Ce calibrage de caméra est particulièrement important lorsque l'on doit obtenir, à partir des images acquises, des informations métriques en vue d'applications de mesures dimensionnelles. Pour obtenir des mesures dimensionnelles précises, il est indispensable de prendre en compte les distorsions géométriques induites par le système optique utilisé, vitesse, etc…) et à limiter l'impact de la distorsion sur ces informations.

Cette étape de calibrage constitue le point initial pour plusieurs applications de la vision artificielle, comme par exemple la reconnaissance et la localisation d'objets, le contrôle dimensionnel de pièces, la reconstruction de l'environnement pour la navigation d'un robot mobile, etc.

Modèle de caméra :

Le modèle de caméra le plus souvent utilisé est le modèle sténopé, dit pinhole.

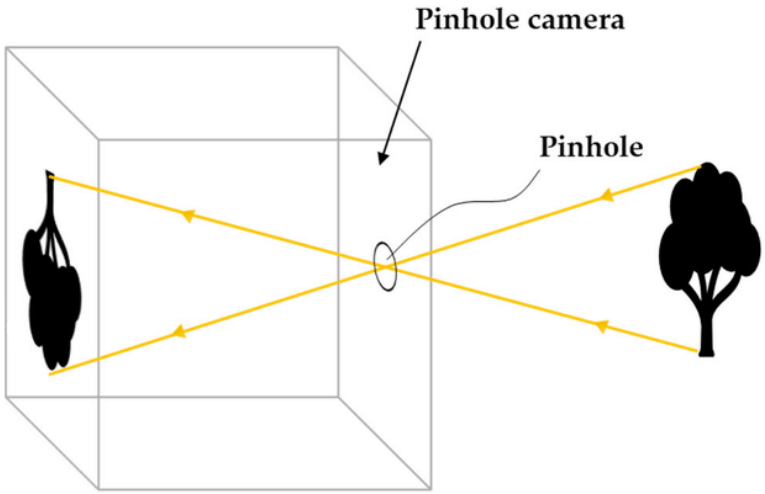


Figure 1: Caméra sténopé (pinhole model)

Le modèle de caméra pin-hole est un modèle dit idéal car il ne tient pas compte de la distorsion de l'objectif. Il est donc souvent utilisé afin de ne pas prendre en compte la distorsion.

Méthode de calibrage :

Pour effectuer un calibrage il existe diverse méthode telle que l'utilisation d'une mire de calibrage.

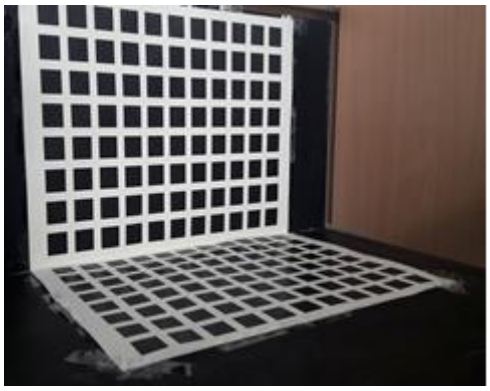


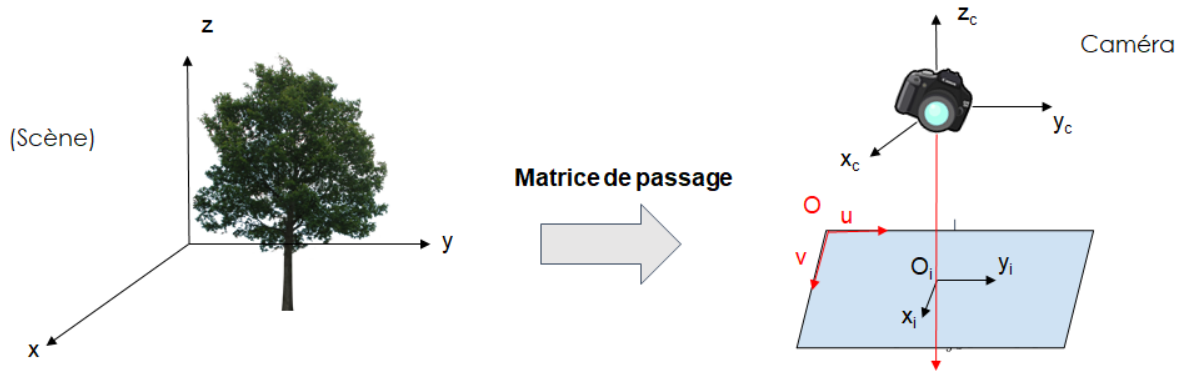
Figure 2: mire de calibrage

Au fil des années, ces méthodes sont devenues de plus en plus sophistiquées pour conduire à un calibrage de plus en plus précis.

Calibrage de caméra :

Nous avons vu précédemment que lors de l'étalonnage, différents paramètres sont nécessaires : les paramètres intrinsèques et extrinsèques.

Les paramètres extrinsèque et intrinsèque de la caméra composent à eux deux la matrice de projection de la caméra. C'est grâce à celle-ci que l'on peut passer des coordonnées 2D associées à l'image et les coordonnées 3D d'un objet du repère réel.



Repère 3D :

(O,x,y,z) : Scène
(C,x_c,y_c,z_c) : Caméra

Repère image:

(O_i,x_i,y_i)
(O_x,u,v)

Matrice extrinsèque :

La matrice extrinsèque décrit l'emplacement de la caméra dans le monde et la direction dans laquelle elle est dirigée. Ils sont liés à la rotation et la translation de la caméra entre le repère de coordonnées réelles et le repère 3D de la caméra.

$$\begin{pmatrix} x_c \\ y_c \\ z_c \\ 1 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$

xc,yc,zc : coordonnées 3d caméra

x,y,z : coordonnées 3ds de la scène (monde réel)

Matrice intrinsèque : ce sont les paramètres de la caméra (lentille, focal, centre optique), ces coefficients étant propre à la caméra ils sont connus.

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \underbrace{\begin{pmatrix} \alpha_x & 0 & u_0 & 0 \\ 0 & \alpha_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}}_{\text{Matrice intrinsèque}} \begin{pmatrix} z_c u \\ z_c v \\ zc \\ 1 \end{pmatrix}$$

- *(x, y, z) : pour les coordonnées réelles*

• *(xc, yc, zc) : les coordonnées du plan de la caméra*

• *(u,v) : Les coordonnées 2D de la caméra*

La matrice intrinsèque est un tableau qui contient les différentes caractéristiques de la caméra qui vont nous permettre d'effectuer la transition entre l'image 3D perçue par la caméra et l'image 2D qu'elle affiche.

La matrice extrinsèque donnée par la position de la caméra, définit le plan de l'image.

A partir du produit de ces deux matrices, nous obtenons une équation nous permettant de passer du réel à l'écran de la caméra.

$$\begin{bmatrix} uz_c \\ vz_c \\ z_c \end{bmatrix} = \begin{bmatrix} \alpha_x & 0 & u_0 & 0 \\ 0 & \alpha_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

À partir de là, on pourra ensuite isoler les vecteurs u et v qui représente le plan vu par la caméra.

Pour cela on commence dans un premier temps à multiplier les matrices :

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{32} & m_{32} & m_{33} & m_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

Où on a pour chaque terme :

$$
\begin{aligned}
m_{11} &= \alpha_x r_{11} + u_0 r_{31} \\
m_{12} &= \alpha_x r_{12} + u_0 r_{32} \\
m_{13} &= \alpha_x r_{13} + u_0 r_{33} \\
m_{14} &= \alpha_x r_{14} + u_0 r_{34} \\
m_{21} &= \alpha_y r_{21} + v_0 r_{31} \\
m_{22} &= \alpha_y r_{22} + v_0 r_{32} \\
m_{23} &= \alpha_y r_{23} + v_0 r_{33} \\
m_{24} &= \alpha_y r_{24} + v_0 r_{34} \\
m_{31} &= r_{31} \\
m_{32} &= r_{32} \\
m_{33} &= r_{33} \\
m_{34} &= r_{34}
\end{aligned}
$$

En isolant les vecteurs u et v on obtient :

$$u = \frac{m_{11}x + m_{12}y + m_{13}z + m_{14}}{m_{31}x + m_{32}y + m_{33}z + m_{34}}$$

$$v = \frac{m_{21}x + m_{22}y + m_{23}z + m_{24}}{m_{31}x + m_{32}y + m_{33}z + m_{34}}$$

Puis en réorganisant :

$$u * (m_{31}x + m_{32}y + m_{33}z + m_{34}) = m_{11}x + m_{12}y + m_{13}z + m_{14}$$
$$v * (m_{31}x + m_{32}y + m_{33}z + m_{34}) = m_{21}x + m_{22}y + m_{23}z + m_{24}$$

On obtient deux équations pour le moment.

L'objectif pour nous va donc être de développer une méthode de calibrage de caméra permettant de déterminer les différents termes m_{xx} afin de déterminer les coordonnée x et y.

Notre méthode nous permet de déterminer la coordonnées de profondeur Z ainsi que les termes de la matrice intrinsèque de la caméra.

L'objectif va donc être de déterminer toutes les inconnues et ainsi de déterminer les coordonnées X et Y que l'on recherche afin que la calibration de la caméra soit complète.

Méthode de calibrage de la caméra :

Comme il a été dit précédemment il existe une multitude de méthode pour pouvoir calibrer des caméras, cela peut être fait en utilisant une mire de calibrage, en utilisant plusieurs images d'une scène sous différents angles ou le mouvement des objets dans la scène. Pour ce stage, l'objectif est de développer une méthode de calibrage automatique et à partir de n'importe quelle image (qu'elles soient amateures ou professionnelles).

Notre méthode de calibration est la suivante :

- Déterminer la profondeur d'un objet par rapport à la caméra
- Déterminer la distance focale et la distorsion d'une image
- Détermination des coordonnées 3D des objets dans une image.

Déterminer la profondeur d'un objet par rapport à la caméra :

Notre méthode de détermination de la profondeur d'un objet repose sur la géométrie des boîtes d'ancrage des objets dans une scène. Il faut donc dans un premier temps déterminer la géométrie de ces boîtes d'ancrage.

La détermination de la profondeur d'un objet s'effectue donc en deux étapes :

1. Détermination des boîtes d'ancrage et classification des objets sur l'image
2. Estimation de la distance

1. Détermination et classification

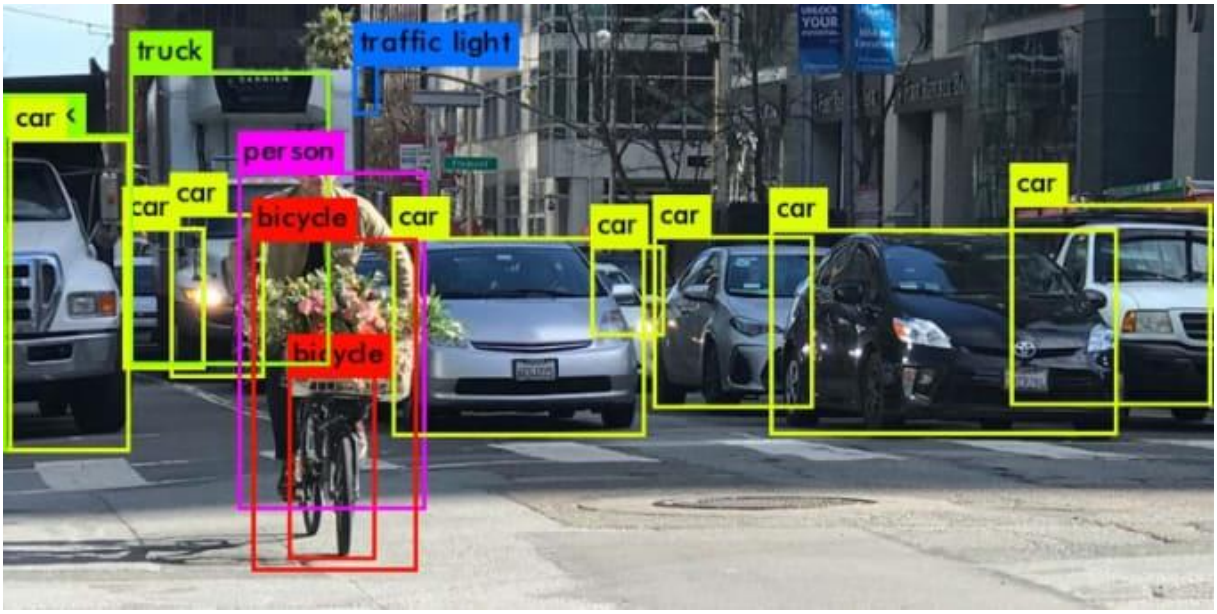


Figure 3: traçage des boîtes d'ancrage et classification des objets sur une image

Pour résoudre ce problème notre choix s'est porté sur le système de détection d'objets YOLO [9].

Nous avons choisi YOLO pour le stage car il est connu, efficace, peut être utilisé dans des applications en temps réel comme des vidéos par exemple et qu'il a été utilisé dans l'étude sur l'estimateur de distance DisNet. [2]

Le modèle YOLO que nous utilisons est pré-entrainé sur 80 classes différentes. YOLO prend en entrée une image et redonne en sortie la classe des objets présents sur l'image accompagnés de leur boîte d'ancrage.

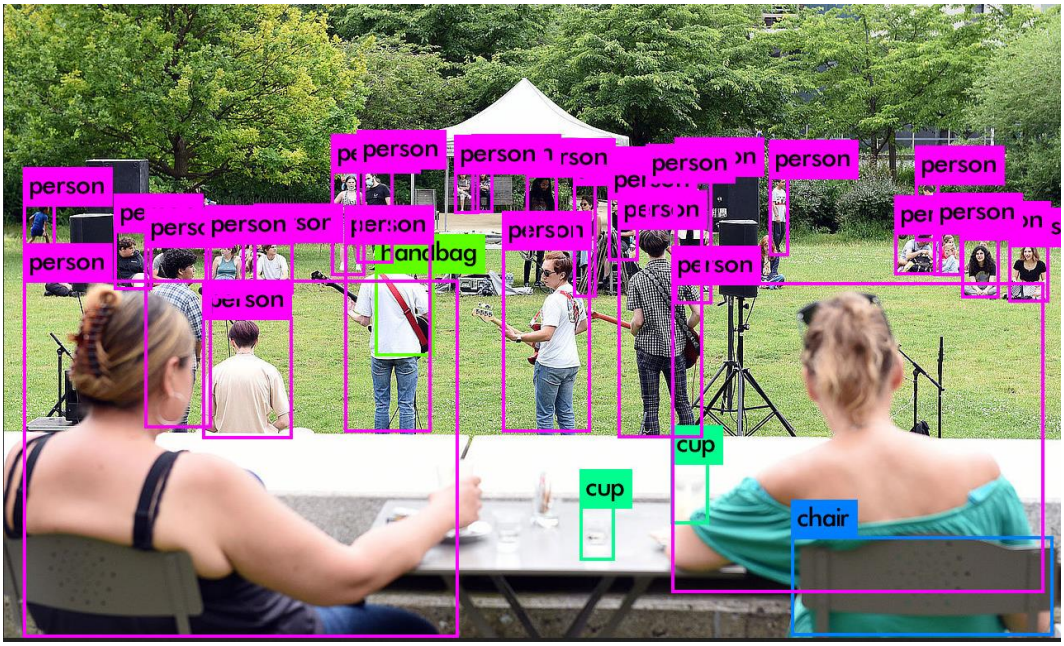


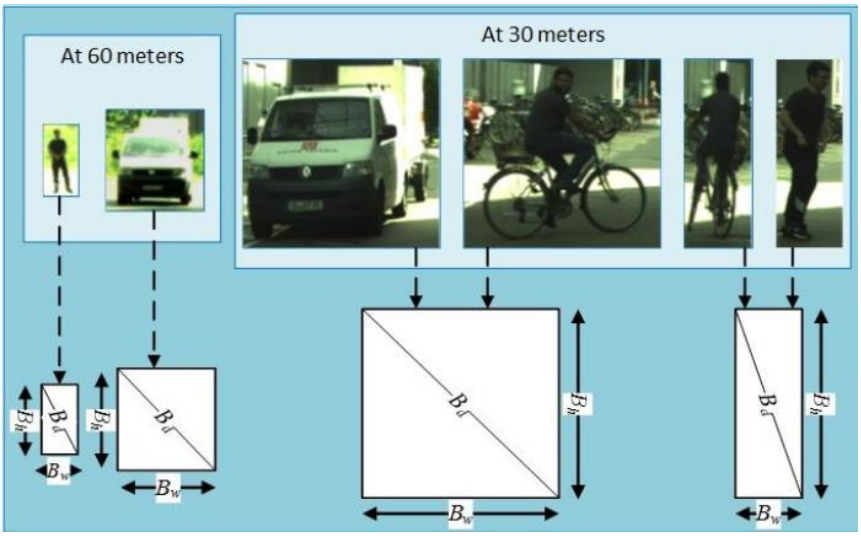
Figure 4 : sortie de yolo

2. Estimation de la profondeur

Notre modèle d'estimation de profondeur s'inspire du travail de Harshil Patel qui lui-même utilise les méthodes d'estimation de profondeur de DisNet [2].

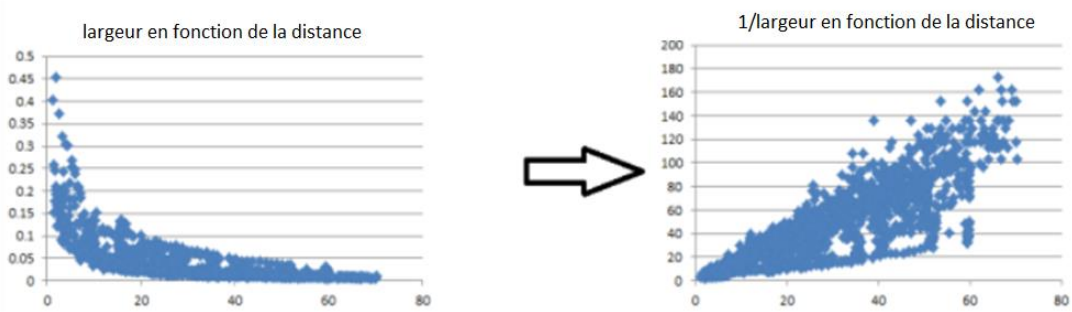
Son modèle d'estimateur de profondeur prend en entrée les géométries des boîtes d'ancrage et donne en sortie la profondeur des objets présents sur l'image.

Le principe utilise le fait que la géométrie des boîtes d'ancrage varie en fonction de la distance de l'objet par rapport à la caméra et que chaque classe d'objets à une valeur de bounding box moyenne qui est propre à sa classe.



En connaissant la relation entre la distance par rapport à géométrie des boîtes d'ancrage (longueur, largeur et diagonale) permet donc d'approximer la position des objets sur la scène.

Il est important de récupérer les valeurs inverses de géométrie des bounding box afin d'avoir une relation linéaire entre la distance est la géométrie de la bounding box.



C'est une méthode ayant de très bons résultats et qui ne nécessite que peu d'informations sur l'image afin de fonctionner, c'est donc un très bon choix de méthode dans le cadre de notre utilisation pour le stage qui est axé , entre autres, sur les vidéos ou images amateures présentes sur le web qui peuvent donc être de mauvaise qualité et qui nécessite une méthode d'estimation de profondeur solide.

En utilisant la géométrie des boîtes d'ancrage et en ayant un modèle de détection et de classification entrainé sur une large base de données (comme YOLO) alors cette méthode peut fonctionner dans de très nombreux cas même dans les cas d'images ou vidéo de mauvaise qualité.

Détermination de la distance focale :

On va ensuite venir déterminer les paramètres de la matrice de passage avec DeepCalib [3]. C'est une approche entièrement automatique basée sur le deep learning qui ne nécessite qu'une image ou une vidéo en entrée.

DeepCalib [3] reçoit en entrée une image et redonne en sortie ses paramètres intrinsèques (distance focale, distorsion).

La connaissance de la distance focale est essentielle afin de pouvoir déterminer les coordonnées X et Y c'est pourquoi nous avons choisi cette méthode.

On va ensuite utiliser ces paramètres, ainsi que la coordonnée Z afin de déterminer avec les matrices précédentes les coordonnées X et Y.