Université Paris 8 - Vincennes à Saint-Denis

Master 1 Informatique

Mémoire projet tuteuré
Qualification de caméras RGB-D

Yasmine BOUDJEMAÏ
Mélanie DE JESUS CORREIA

Organisme : Université Paris 8 Vincennes-Saint-Denis
Tuteur : Farès BELHADJ

Dédicaces

Remerciements

Résumé

Table des matières

Table des figures

Introduction

Introduction a faire à la fin.

Chapitre 1

État de l'art

Dans ce chapitre, nous allons définir ce qu'est une caméra RGB-D, nous énumérons certains des différents modèles existants, nous abordons brièvement les différentes techniques utilisées pour la récupération de la profondeur par une caméra. Enfin, nous discutons les différences recensées sur la `kinect V2` et `V3`.

1.1 Description d'une caméra RGB-D

La caméra RGB-D, aussi appelée capteur RGB-D, est une caméra fournissant en même temps une image couleur et une carte de profondeur caractérisant la distance des objets vus dans l'image. Cela est rendu possible grâce à un capteur RGB et un capteur de profondeur (D pour Depth). C'est principalement ces captures qui vont nous intéresser tout au lond des qualifications.

1.2 Modèles existants

Il existe différents modèles de caméras RGB-D. Parmi elles, nous pouvons citer la Kinect et ses différentes versions, Asus Xtion Pro Live, BlasterX Senz3D, Orbbec, Intel RealSens D415, ...

`La Kinect` a fait son apparition en septembre 2008. Elle a été conçue par Microsoft et était destinée pour la console de jeu XBox 360. Elle permettait aux utilisateurs d'interagir avec la console à l'aide d'une NUI [1] en utilisant les mouvements gestuels et une reconnaissance vocale. Elle sera plus tard

1. Natural User Interface (Interface Utilisateur Naturelle), se réfère à une interface utilisateur invisible.

utilisée dans les domaines de la recherche et du développement pour différents secteurs comme le domaine de la médecine, l'industrie automobile, la robotique, l'éducation, ...

`Asus Xtion Pro Live` est le modèle de référence que nous utilisons afin d'effectuer les qualifications. Elle utilise la technologie PrimeSense [2] pour la détection de mouvements.

`BlasterX Senz3D` a fait son apparation en Septembre 2016. Conçue par Creative, elle a été présentée comme une webcam intelligente basée sur ce qu'il y a de mieux en matière de savoir-faire et de technologie. Elle possède trois lentilles pour capturer les données visuelles : une caméra RVB, une caméra infra-rouge et un projecteur laser. Ces dernières collaborent avec la technologies Inetl RealSense pour réagir aux expressions faciales et aux gestes corporels des utilisateurs.

`Orbbec Astra Pro` fait partie de la série Astra. Elle offre une vision par ordinateur qui permet des dizaines de fonctions telles que la reconnaissance des visages, la reconnaissance des gestes, le suivi du corps humain, la mesure tridimensionnelle, la perception de l'environnement et la reconstruction de cartes en trois dimensions. De plus, elle offre une réactivité haut de gamme, une mesure de la profondeur, des dégradés fluides et des contours précis ainsi que la possibilité de filtrer les pixels de profondeur de faible qualité.

`Intel RealSense D415` a un champ de vision standard bien adapté aux applications de haute précision telle que la numérisation 3D. Elle comprend le processeur Intel RealSense Vision D4 offrant une résolution en profondeur élevée, des capacités de longue portée, une technologie d'obturation globale et un large champ de vision. Grâce à ces deux dernières, la caméra offre une perception précise de la profondeur lorsque l'objet est en mouvement ou que l'appareil est en marche. De plus, elle couvre un champ de vision plus large, minimisant ainsi les angles morts.

1.3 Erreur de mesure

L'erreur de mesure, autrefois appelée l'erreur absolue, est un processus qui permet d'évaluer l'écart entre la valeur mesurée et la valeur de référence qui est soit exacte ou connue. L'objectif de la mesure d'erreurs est de jauger à quel degré les deux valeurs sont proches. Dans cette section, nous citons les types ainsi que quelques sources d'erreurs, ensuite, nous faisons un détour

2. Connu principalement pour sa licence de conception matérielle et de puce employée dans le mécanisme de détection de mouvements de la Kinect XBox360. Pour plus d'informations, le lecteur peut se référer à ce lien `https://www.crunchbase.com/organization/primesense#section-web-traffic-by-similarweb`.

sur les moyens théoriques et physiques employés à cet effet.

1.3.1 Types et sources d'erreurs

Il existe deux types d'erreurs : aléatoires et systématiques.
Les erreurs aléatoires sont des erreurs dont les valeurs sont incohérentes et imprévisibles même en répétant les observations et en conservant les mêmes paramètres.
Les erreurs dites systématiques, à l'inverse des erreurs aléatoires, sont des erreurs reproductibles, elles demeurent constantes tant que les paramètres restent inchangés.

Ces erreurs peuvent résulter de différentes sources ; elles peuvent être dues à des facteurs environnementaux, à la résolution de l'instrument utilisé, au calibrage de l'appareil employé ou même à des erreurs humaines.

1.3.2 Mesure d'erreurs à l'aide de matériels

1.3.3 Mesure d'erreurs par une approche théorique

A ce jour, on recense de multiples méthodes exploités à cet effet. Nous nous contentons de citer ci-dessous celles qui nous paraissent êtres les plus pertinentes.

RMSE/RMSD : Root Mean Square Error ou encore Root Mean Square Deviation que l'on pourrait traduire par ; la racine carré de la moyenne des erreurs au carré. Elle demeure parmi les plus fréquentes en la matière. C'est une mesure de l'exactitude, qui permet de comparer les erreurs de prévision de différents modèles pour un ensemble de données précis et non entre des ensembles de données. Sa formule s'exprime comme suit :

$$\sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \hat{x_i})^2} \tag{1.1}$$

De par la formule, on peut déduire que le résultat sera toujours positif et qu'obtenir une valeur de 0 (qui n'est généralement non atteint en pratique), témoignerait d'une exactitude parfaite. Ainsi, obtenir des valeurs avoisinant le zéro est positif pour les mesures.

MSE : Mean Square Error que l'on pourrait traduire par ; l'erreur quadratique moyenne. Sa formule est comme suit :

$$\frac{1}{N}.\sum_{i=1}^{N}(x_i - \hat{x_i})^2 \tag{1.2}$$

MAE : Mean Absolute Error qui pourrait être traduit par ; erreur moyenne absolue. Son équation est comme ci-dessous :

$$\frac{1}{N}.\sum_{i=1}^{N}|x_i - \hat{x_i}| \tag{1.3}$$

MAPE : Mean Absolute Percentage Error qui pourrait être traduit par ; le pourcentage de l'erreur moyenne absolue. Elle fait partie également des moyens de mesures d'erreur les plus fréquents. A noter que la valeur observée ne doit être aucunement nulle. La formule est comme ci-après :

$$\frac{1}{N}.\sum_{i=1}^{N}\left|\frac{x_i - \hat{x_i}}{x_i}\right| \tag{1.4}$$

Avec :

x_i, la valeur observée à la ième observation.
$\hat{x_i}$, la valeur exacte à la ième observation.
N, représente le nombre d'observations

1.4 Technologies utilisées pour la depth d'une caméra RGB-D

Des méthodes employées pour la récupération de la depth, on peut en citer deux principales.

1.4.1 Infrarouge

Le projecteur infrarouge de certaines caméras le possédant qu'on peut voir sur la figure 1.1 projette un spectre infrarouge sur la scène captée. Le motif produit sur cette dernière sera capté par la caméra infrarouge et sera par la suite comparé à la base de données de motifs de référence stockés au préalable dans la caméra. Ces derniers seront indispensables pour la mesure de la profondeur de chaque pixel.

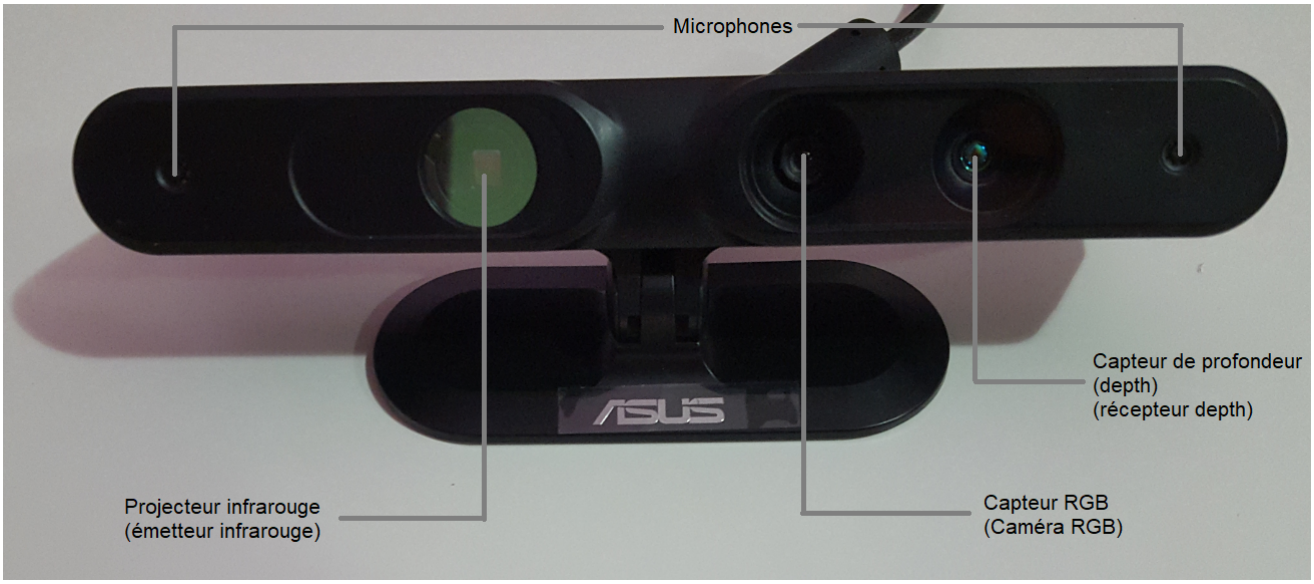

FIGURE 1.1 – Composants caméra RGB-D (Ici Asus Xtion Pro Live).

Par la suite, les valeurs obtenues seront corrélées à un capteur RGB qu'on peut apercevoir sur la figure. Ces données pourront être représentées par un nuage de points[3].

Toutefois, ce type de caméras possède certaines restrictions. Parmi ces dernières, on peut citer :

— Distances de mesure limitées.
— Problèmes de calculs des informations de profondeur à l'encontre de surfaces brillantes, très mates, transparentes, réfléchissantes ou encore envers des objets absorbants.
— Interférence des motifs (patterns) infrarouges si présence de plusieurs caméras RGB-D de même type. En effet, chaque capteur visualisera ses propres motifs ainsi que ceux des autres caméras présentes et ne saura distinguer les siens des autres qui se chevauchent. De cela découle une perte considérable d'informations de profondeur comme nous pouvons l'observer sur la figure 1.2.
Cependant, des travaux de recherches ont été conduits afin d'y remédier. Parmi eux, on peut citer le travail réalisé par l'équipe de *Rafibakhsh* ([**RAFIBAKHSH15**]) qui recommande de laisser un angle de 35° entre deux caméras suspendues à la même hauteur en considé-

3. Est une représentation des points de coordonnées tridimensionnelles dont chacun peut avoir des attributs qui lui ont sont propres.

FIGURE 1.2 – (De gauche à droite) image renvoyée par la caméra RGB, image IR, depth map (une seule caméra), depth map (deux caméras). (figure tirée de l'article de *F.Alhwarin, A.Ferrein* et *I.Scholl* ayant comme titre *IR Stereo Kinect : Improving Depth Images by Combining StructuredLight with IR Stereo*).

rant les scènes captées dans de bonnes conditions et une interférence très faible. *Maimone* et *Fuchs* [**MaimoneFuchs15**] proposent un algorithme de remplissage et de lissage en modifiant le filtre médian aux zones trouées à l'exception des bords. Quant à *F. Kenton Musgrave, Craig E.* et *Robert S. Mace* [**KentonCraigMace12**], ils appliquent une certaine quantité minime de mouvements (en utilisant des composants matériels supplémentaires) à certains capteurs de sorte que chacun puisse voir son propre motif infrarouge de façon nette et une version floue des motifs de ses voisins.

1.4.2 Stéréo-vision

Une caméra stéréoscopique est un appareil qui contient deux voire plus de capteurs d'images. Ceci nous rappelle la vision binoculaire humaine. En effet, ce mécanisme permet au système nerveux central de percevoir simultanément les images issus de chaque œil envoyées sous forme de signaux. Ainsi, il sera en mesure de se servir de ces différences (entre les deux images) pour permettre une vision stéréoscopique pour la perception de relief et une mesure des distances en utilisant la triangulation [4]. Le concept que nous venons d'expliquer est appelé disparité stéréoscopique [5].

La stéréo-vision sert principalement à reconstituer la scène observée sous forme de modèle 3D.

4. Approche géométrique permettant une mesure des distances. Le lecteur peut consulter cette page web pour de plus amples informations : `https://fr.wikipedia.org/wiki/Triangulation` .

5. Différence dans la localisation d'un objet perçu par l'œil gauche et l'œil droit résultant de la séparation horizontale des yeux dont certaines caméras essaient de s'approprier la technique afin de récupérer la profondeur.

Carte de disparités

Le principe de la disparité est une approche du mécanisme humain vu précédemment. Il consiste en la différence entre les coordonnées pixels d'un point bidimensionnel d'une image et celles de son correspondant (présent sur une autre image prise au même moment). Ainsi, en appliquant le même traitement sur tous les pixels correspondants, on obtient la carte des disparités. Une carte est dite éparse lorsqu'une disparité est associée à quelques pixels et lorsque cette dernière est associée à chaque pixel, on dit d'elle qu'elle est dense.
Une formule pour calculer la profondeur en fonction de la disparité s'obtient comme suit :

$$z = \frac{B.f}{d} \tag{1.5}$$

Avec :

z représentant la profondeur.
B *Baseline* représente la distance séparant les deux capteurs.
f La distance focale en pixels calculée comme montré ci-après en nous servant de la figure 1.3 .

$$\tan\left(\frac{fov}{2}\right) = \frac{D}{2.f} \Leftrightarrow f = \frac{D}{2.\tan\left(\frac{fov}{2}\right)} \tag{1.6}$$

Avec :

D dimension qui peut représenter soit la largeur soit la hauteur en fonction du fov de la caméra employé.
fov (Field Of View) représente l'angle du champ de vision (soit horizontal ou vertical).
f représente la distance focale.

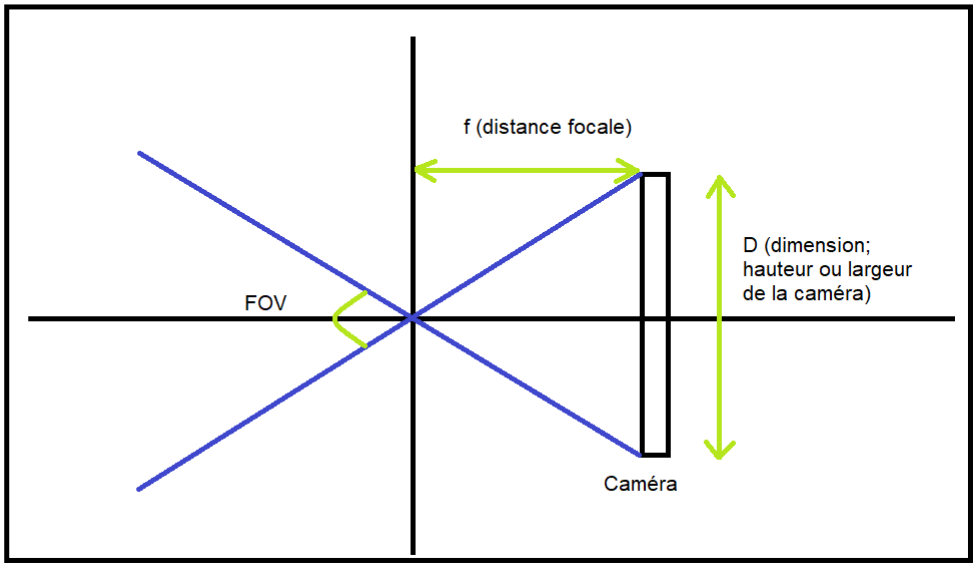

FIGURE 1.3 – Schéma illustrant la relation entre la distance focale et le champ de vision (fov).

1.5 Kinect V2 et Kinect V3

1.5.1 Kinect V2

`Kinect V2` se sert de la méthode TOF (ie : Time Of Flight ou autrement dit Temps De Vol) pour générer la carte de profondeur. Cette technique se base sur la différence de temps entre l'émission d'un faisceau lumineux et son retour après réflexion sur un objet.
La distance est calculée comme suit :
$d = c.\frac{\Delta t}{2}$, avec c la vitesse de la lumière dans l'air.

Elle permet une bien meilleure précision même dans le noir que sa version précédente (`Kinect V1`).

1.5.2 Kinect V3

`Kinect V3`, baptisée `Microsoft Azure Kinect V3`, tout comme la précédente version, se base aussi sur la technologie TOF. Elle comprend un capteur RGB de 12 Mp, un capteur de depth de 1Mp avec un fov réglable en large ou réduit ainsi que 7 microphones intégrés.

Kinect V2 Vs Kinect V3

`Kinect V3` est beaucoup plus légère et plus petite que la `V2`. Elle possède entre autre plus de microphones que la précédente. Elle a été conçue afin d'être principalement utilisée avec Azure , le service cloud de `Microsoft`. C'est sans doute ce qui la démarque le plus de sa prédécesseure. En effet, ce service lui permet d'effectuer une partie des calculs. En outre, elle bénéficie des Cognitive Services, autrement dit de l'intelligence artificielle pourra être incluse dans les applications créées.

Chapitre 2

Notre outil pour la qualification de caméras RGB-D

Dans ce passage, nous allons décrire notre outil conçu pour qualifier les caméras par rapport à notre modèle-étalon montré dans l'image ci-dessous.

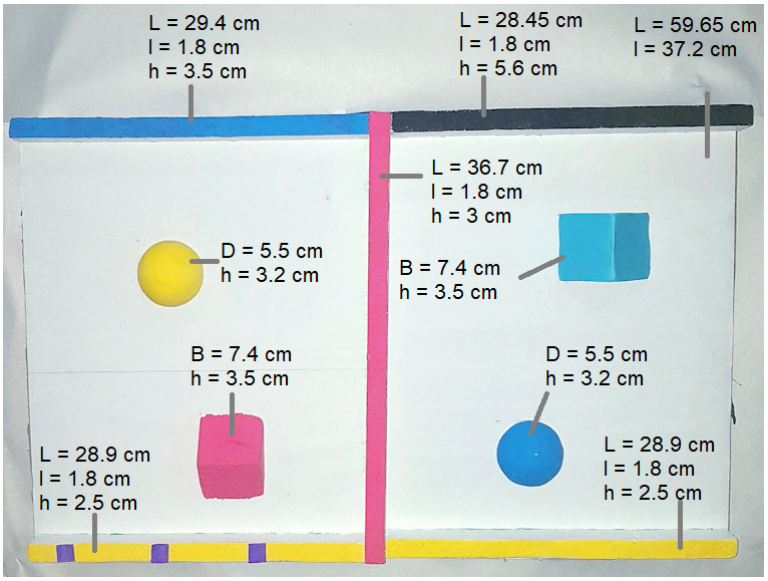

FIGURE 2.1 – Modèle-étalon.

Chapitre 3

Modèle logiciel pour la qualification de caméra RGB-D

Dans ce passage, nous allons décrire notre outil conçu pour qualifier les caméras.

3.1 Description de l'application

3.1.1 Première partie : Affichage de la scène captée

3.1.2 Seconde partie : Interface graphique

3.2 Conception du modèle réel et du modèle virtuel

3.3 Comparaison de la depth OpenGL avec la depth de la caméra RGB-D

Dans ce passage, nous exposons certains des résultats obtenus ainsi que les résultats attendus avec la caméra Asus.

3.3.1 Résultats obtenus

3.3.2 Résultats attendus

Chapitre 4

Cas pratiques de qualification

4.1 Modèle 3D

4.2 La POC

4.3 Résultats et critique

Chapitre 5

Conclusion et Perspectives