IDENTIFICATION DE LA ZONE REGARDÉE SUR UN ÉCRAN D'ORDINATEUR À PARTIR DU FLOU

par

Eric Néron

Mémoire présenté au Département d'informatique
en vue de l'obtention du grade de maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES

UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, 29 novembre 2017

Le 29 novembre 2017

le jury a accepté le mémoire de Monsieur Eric Néron
dans sa version finale.

Membres du jury

Professeur Djemel Ziou
Directeur de recherche
Département d'informatique

Professeur Reza Jafari
Évaluateur externe
Institut Optina Diagnostics

Professeur Marie-Flavie Auclair-Fortier
Président rapporteur
Département d'informatique

Sommaire

Quand vient le temps de comprendre le comportement d'une personne, le regard est une source d'information importante. L'analyse des comportements des consommateurs, des criminels, ou encore de certains états cognitifs passe par l'interprétation du regard dans une scène à travers le temps. Il existe un besoin réel d'identification de la zone regardée sur un écran ou tout autre médium par un utilisateur. Pour cela, la vision humaine fait la composition de plusieurs images pour permettre de comprendre la relation tridimensionnelle qui existe entre les objets et la scène. La perception 3D d'une scène réelle passe alors à travers plusieurs images. Mais qu'en est-il lorsqu'il n'y a qu'une seule image ? Dans ce mémoire, pour comprendre cette relation nous allons discuter de la formation d'image dans son ensemble. Par la suite, pour tenter de comprendre le lien entre la profondeur et la perception d'un monde en 3D, vont suivre les concepts et la terminologie concernant le phénomène que représente le flou à travers une image. Cette théorie nous permet de construire deux modèles d'estimation du regard dont nous allons évaluer les résultats. La différence entre les modèles se situe dans la façon d'inférer l'information 3D qui est utilisée pour localiser la zone de la scène regardée.

Sommaire

Remerciements

Je souhaite remercier quelques personnes m'ayant encouragé tout au long de cette aventure que fut la maîtrise. En commençant par le laboratoire MOIVRE et les collègues s'y trouvant qui ont été inspirants par les conversations et le partage d'idées. Tout particulièrement le doctorant Julien Couillaud qui a été une aide remarquable en plus de servir de référence tout long de mes années au MOIVRE. Également, un remerciement spécial à mon directeur de recherche Djemel Ziou, qui m'a dirigé et guidé tout au long du processus en plus de parfaire mes connaissances et ma culture dans mon domaine de recherche. Je remercie les participants qui m'ont permis de construire une base de données relative aux besoins de mon système d'estimation du regard. Finalement, je remercie mes amis, ma famille et le jury pour la lecture et l'évaluation de ce mémoire.

Remerciements

Épigraphe

"Si t'as un bébé, je te donne 100$
pis t'en fais un capitaine."

Gab le Marin

Épigraphe

Abréviations

CCD *Charge-coupled device*

CCHST Centre canadien d'hygiène et de sécurité au travail

CdC Cercle de confusion

CFA *Color filter array*

CMOS *Complementary metal-oxide-semiconductor*

DoF *Depth of field*

E Irradiance

IR Infrarouge

IRM Imagerie par résonance magnétique

K Kilojoule

L Radiance

ICIAR *International Conference on Image Analysis and Recognition*

MAP Maximum a posteriori

MV Maximum de vraisemblance

PDS Puissance de distribution spectrale

PSF *Point spread function*

UV Ultraviolet

VBMLR *variational Bayesian multinomial logistic regression*

RO Région observée

Table des matières

Sommaire iii

Remerciements v

Épigraphe vii

Abréviations ix

Table des matières xi

Liste des figures xv

Liste des tableaux xvii

Introduction 1

1 Formation d'image 5

1.1 Introduction 5

1.2 Formation d'image 6

1.3 Systèmes de formation passif de l'image 7

1.3.1 Système actif 7

1.3.2 Système passif 8

1.4 Source lumineuse 10

1.5 Introduction à la notion de réflectance par la scène 12

1.6 Formation géométrique de l'image 15

1.6.1 Repère de la scène et repère caméra 15

1.6.2 Modèle de caméra à sténopé 17
1.7 Formation optique de l'image 19
1.7.1 Modèle de caméra à lentille mince 19
1.8 Formation radiométrique de l'image 21
1.9 Traitement physique de l'image (capteur) 25
1.10 Conclusion 28

2 Concepts et terminologies sur le flou **29**
2.1 Introduction 29
2.2 Définition du flou 30
2.2.1 Flou constant 30
2.3 Cercle de confusion (CdC) 30
2.4 Profondeur de champ (DoF) 33
2.5 Fonction d'étalement d'un point (PSF) 36
2.5.1 Le patron d'Airy 36
2.5.2 Modèle gaussien 38
2.6 Conclusion 40

3 Estimation du regard **41**
3.1 Introduction 42
3.2 System overview 43
3.3 Mapping estimation 44
3.4 Feature space 47
3.4.1 Iris disparities estimation 48
3.4.2 Head position and orientation using a marker 50
3.4.3 Head position and orientation using the blur 53
3.5 Experiments 55
3.6 Conclusion 59

4 Résultats expérimentaux **65**
4.1 Introduction 65
4.2 Profondeur z avec marqueur 66
4.3 Profondeur z selon σ 70

4.4 Estimation du regard : Expérimentation 1 74
4.5 Estimation du regard : Expérimentation 2 77
4.6 Estimation du regard : Expérimentation 3 81
4.7 Estimation du regard : Temps de calcul des performances 83
4.8 Conclusion 84

Conclusion **87**

Liste des figures

1.1 Système actif 8
1.2 Système passif 9
1.3 Spectre lumineux, image provenant de [36] 10
1.4 Graphique représentant l'évolution de la densité spectrale d'énergie émise par un corps noir en fonction de la longueur d'onde pour plusieurs températures 11
1.5 Interactions lumière/surface 12
1.6 Réfraction sur une surface 13
1.7 Modèle de réflexion spéculaire (Phong et Blinn) 14
1.8 Représentation des deux repères orthonormés 16
1.9 Modèle de caméra à sténopé 17
1.10 Modèle de caméra à lentille mince 20
1.11 Modèle de caméra à lentille mince avec zones de flou 21
1.12 Modèle radiométrique avec angle solide 23
1.13 Quantité de lumière absorbée par la lentille 24
1.14 Exemple de capteur CCD 26
1.15 Interpolation par bloc 27

2.1 Représentation d'un cercle de confusion 32
2.2 Zone de la profondeur de champ 33
2.3 Profondeur de champ selon l'ouverture 34
2.4 Calcul de la profondeur de champ 35
2.5 Courbes de niveaux pour le Patron d'Airy et le modèle gaussien, provenant de [39] 38

2.6 Relation géométrique entre σ et z, provenant de [45] 39

3.1 Gaze estimation system framework 45
3.2 Marker and nose tip detection example 48
3.3 Iris detection example 49
3.4 Iris detection for particular cases 49
3.5 Marker coordinates system between the scene and the image plane . . 51
3.6 Nose tip and head center detection 55
3.7 Representation of the reference point $T(x, y, z)$ and the nose tip $N(x, y, z)$ in the scene 56
3.8 17 inches screen subdivision 57
3.9 Nose detection 58

4.1 Représentation de z par rapport aux images acquises pour 3 sujets. Le tracé rouge représente la moyenne des 3 sujets 70
4.2 Courbe de z selon σ 71
4.3 Profondeur Z selon σ 71
4.4 Représentation de σ par rapport aux images acquises pour trois sujets. Le tracé rouge représente la moyenne des 3 sujets 73
4.5 Environnement de suivi du regard lors l'expérimentation 1 74
4.6 Environnement de suivi du regard lors des expérimentations 2 et 3 . . 78

Liste des tableaux

3.1 Training = 250, Testing samples (#) $\leq$ 50 59
3.2 Training = 250, Testing samples (#) $\leq$ 50 59
3.3 Training = 600, Testing samples (#) $\leq$ 75 60

4.1 Résultats en % de la classification de l'erreur pour 500 images à une distance X. 67
4.2 Évaluation de z en mm pour 3 sujets à travers les images acquises . . 68
4.3 Évaluation de σ à travers les images acquises 73
4.4 Apprentissage = 100 données, tests $\leq$ 50 données 76
4.5 Apprentissage = 250 données, tests $\leq$ 50 données 76
4.6 Apprentissage = 100 données, tests $\leq$ 50 données 77
4.7 Apprentissage = 250 données, tests $\leq$ 50 données 77
4.8 Sujet 1 : apprentissage = 250 données, test $\leq$ 50 données 78
4.9 Sujet 2 : apprentissage = 250 données, test $\leq$ 50 données 79
4.10 Sujet 1 : apprentissage = 250 données, tests $\leq$ 50 données 79
4.11 Sujet 2 : apprentissage = 250 données, tests $\leq$ 50 données 80
4.12 apprentissage = 600 données, tests = 75 données 82
4.13 apprentissage = 450 données, tests = 75 données 83
4.14 Évaluation des performances des différentes étapes de l'application . . 84

Introduction

L'essor des technologies dans les années 90 a mené à de nombreuses applications dans le domaine de la vision par ordinateur. Que ce soit en cinéma, en robotique, en médecine ou encore en aérospatiale, le traitement d'image devient un incontournable dans le développement de ces nouvelles applications [7]. Avec des technologies de plus en plus prometteuses, la recherche constante d'une meilleure qualité d'image et une plus grande rapidité calculatoire font foi de qualité des applications développées. Bien que visuellement, pour un humain, la qualité d'une image soit excellente, certains systèmes de vision demandent une qualité encore plus grande pour performer adéquatement [18, 37]. Cependant, en plus de devoir nous conformer à un courant technologique de performance, nous sommes dans une ère de développement compact. C'est-à-dire que nous réduisons le médium au maximum afin de le transporter à nos côtés. Le téléphone, l'appareil photo, l'agenda, les cartes de paiements, tout doit être pensé pour la facilité et la mobilité. Malheureusement, cela conduit à une qualité d'image limitée par la performance des technologies qui sont à notre portée comme les caméras intégrées dans les appareils mobiles et les ordinateurs portables. Avec l'utilisation de caméras, plusieurs caractéristiques liées à la tridimensionnalité de notre monde réel doivent être rendues par l'entremise de capteurs qui eux, transforment la lumière captée par des cavités photosensibles en signal électrique, pour former une image en deux dimensions. Malheureusement, la miniaturisation de plusieurs composantes affectent significativement la qualité de l'image. Par exemple, la dimension des capteurs qui sont de plus en plus petits de même que l'ouverture de la caméra sont les principaux éléments qui affectent la qualité d'une image. Cela conduit à intégrer des algorithmes de traitement d'images qui doivent performer sous des conditions loin d'être idéales pour compenser la qualité. La recherche d'algorithme dans le domaine

de l'imagerie n'a donc jamais été autant d'actualité.

Dans un système de formation d'image, en utilisant un plan en deux dimensions pour former une image, nous perdons de l'information sur la luminosité de la scène et cette information est parfois primordiale pour travailler l'image par rapport à une scène et en extraire les caractéristiques souhaitées pour certaines applications. Pour récupérer l'information de profondeur, nous devons remonter à l'intérieur du système de formation d'image et comprendre les différentes parties du système qui influencent l'information recueillie par le capteur. La première partie du système de formation d'image, si l'on suit la direction d'un faisceau lumineux cheminant de la scène vers le capteur est l'optique, qui sert à focaliser la lumière. Ensuite, il y a la matrice de filtres de couleurs qui fait l'échantillonnage du spectre de la lumière [5, 4] et qui est une partie intégrante du capteur. Pour sa part, le capteur intègre la lumière et la transforme en signal électrique pour former une image numérique [33, 28]. Toutes ces composantes à l'intérieur d'un système améliorent ou détériorent l'information de la radiométrie ou de la géométrie d'une image. Nous allons donc, dans le chapitre 1 de ce mémoire, examiner en détail les étapes de la formation d'image pour un système de caméra optique.

Le système de caméra optique est le plus utilisé dans la vie courante bien que différents types de systèmes existent. Il y a par exemple, des systèmes de formation d'image conçus pour des situations précises ou pour une cueillette d'information qui diffère de la géométrie de la scène. Comme les systèmes par résonances magnétiques qui permettent d'obtenir des images sur la structure des différents organes du corps [19]. En utilisant un système de caméra optique, nous obtenons une image en deux dimensions. Certaines caractéristiques qui sont reproduites dans l'image permettent de voir la tridimensionnalité de la scène. Ces caractéristiques sont ce qu'on appelle des indices de profondeurs monoculaires ; la texture, l'ombrage, la perspective et le flou. L'optique entraîne donc un flou causé par la lentille. Ce phénomène que nous allons étudier dans le chapitre 2 peut paraitre à première vue superflu, mais nous allons voir par ses concepts et sa terminologie qu'il est essentiel pour la compréhension du processus de formation d'image et aussi très utile si nous arrivons

à bien cerner l'information qu'il peut apporter. Nous allons donc définir le flou et l'observer à travers les cercles de confusion, la profondeur de champ et les fonctions d'étalement d'un point en plus d'établir des liens directs entre le flou et la profondeur.

Dans le chapitre 3, nous allons proposer deux méthodes d'estimation du regard. Dans le domaine de la vision par ordinateur, ce type d'application en temps réel est très demandée pour des champs aussi divers que le marketing [38, 27], la psychologie [35, 34, 10] et le milieu de l'automobile [17, 21]. On définit l'estimation du regard comme étant la position 3D que suit la ligne du regard dans le monde réel à travers le temps. Le développement d'une méthode qui fixe le point qu'un utilisateur observe dans une scène correspond à l'objectif principal de cette maîtrise. On fixe également comme exigence que les méthodes fonctionnent par l'entremise d'une seule caméra, ce qui complique relativement la tâche comparativement aux méthodes qui utilisent la stéréovision ou plusieurs caméras [1, 2, 32, 43]. Nous nous limitons à l'utilisation d'une webcam intégrée dans un ordinateur. Plusieurs méthodes qui utilisent une seule caméra ont également été proposées [13, 24, 25], mais il reste très difficile d'avoir une estimation précise du regard. L'information de profondeur récupérée constitue l'élément central des deux méthodes. Pour le calcul de la profondeur, une méthode utilise un marqueur sur l'individu et l'autre méthode utilise seulement l'information du flou dans les images d'acquisitions.

Chapitre 1

Formation d'image

1.1 Introduction

On peut introduire la formation d'image par deux questions posées par Berthold K.P. Horn en 1986 dans [16] et qui demeurent aujourd'hui fondamentales :

1. Qu'est-ce qui détermine l'emplacement d'un point d'un objet sur l'image ?
2. Qu'est-ce qui détermine la luminosité de la surface d'un objet sur l'image ?

Ce sont deux questions qui nous guideront pour la suite du chapitre et vu les similitudes qui se trouvent entre le système visuel humain et un système de formation d'image [29], il sera facile de tracer plusieurs parallèles entre les deux pour expliquer l'entièreté d'un système de formation d'image. D'abord, nous avons besoin de trouver la transformation géométrique correspondante à un point de la scène à celui d'un point de l'image. Puis, il nous faut trouver comment quantifier la luminosité d'un point particulier de la scène vers l'image. Pour répondre à ces questions, il est important d'avoir une vue d'ensemble sur ce que représente un système de formation d'image. C'est pourquoi nous allons commencer ce chapitre en définissant la formation d'image et la perception d'une scène, à travers un sytème de formation. Ensuite, nous allons poursuivre avec les questions posées par Horn en parcourant le système de formation d'image et ces concepts en entier.

1.2 Formation d'image

Le système visuel humain est par définition l'ensemble des organes participant à la perception visuelle, de la rétine jusqu'au système sensori-moteur. Il a comme rôle de percevoir et d'interpréter deux images en deux dimensions, une image pour chaque oeil, en une image en trois dimensions qui reflète le monde dans lequel on vit. L'humain, pour la grande majorité, possède un système de formation pour chaque oeil et la fusion des deux systèmes forment la perception de la scène. La biologie du système est singulière à chaque personne [23], par conséquent la scène du monde réel qui est capté par notre système visuel est unique à chacun. Avec ses yeux, le système visuel humain possède deux systèmes dépendants de formation d'image. On résume un système de formation d'image comme étant la perception de l'information de la scène qui elle, peut paraître différente pour chacun, à cause de l'analyse que notre système visuel en fait. L'objectif d'un système de formation d'image est donc de réussir à uniformiser un lieu, un endroit, ou un objet.

Une image est le résultat du cheminement de la lumière, provenant d'une source lumineuse, voyageant à travers une scène tridimensionnelle qui parcourt un système de formation d'image pour aboutir sur un récepteur : la rétine pour le système visuel humain, un capteur ou un film pour un système informatisé ou mécanique. Dans notre cas, nous parlerons toujours de capteur. Celui-ci forme un plan en deux dimensions de l'image lors de la lecture de l'information lumineuse. Que ce soit le système visuel humain ou un système machine le processus demeure principalement le même [29]. Comme nous vivons dans un monde visuel composé de trois dimensions, il est important d'avoir les éléments et les connaissances qui nous permettent de représenter et d'effectuer la réduction de dimension pour aller vers le monde image le plus fidèlement possible, c'est-à-dire avec le moins de perte d'information. L'information recueillie permet ensuite de comprendre les relations qui existent entre les différents objets de l'espace tridimensionnel. Elle varie selon la nature du système de formation d'image et peut être par exemple de l'information géométrique entre les objets de la scène, de l'information concernant l'intensité lumineuse ou même de nature structurelle par rapport à un objet de la scène.

Nous allons maintenant survoler les différents types de systèmes de formation d'image. Par la suite, nous parcourrons en détail le système de formation d'image optique puisque c'est celui que nous allons utiliser pour l'élaboration d'une application d'estimation du regard qui sera présenté dans le chapitre 3.

1.3 Systèmes de formation passif de l'image

Le monde de la vision par ordinateur se divise principalement en deux types de systèmes de formation d'image. On peut retrouver des systèmes de formation dits actifs et des systèmes dits passifs. La distinction entre les deux se situe à la nature de la source éclairante. Cette différence est importante, car l'information acquise qui parcourt le système diffère selon la nature de la source qui peut être autant de la lumière que du son. L'humain possède un système de formation passif, mais il existe plusieurs êtres vivants qui possèdent des systèmes dont le principe est différent. Par exemple, pour leur vision, les chauves-souris utilisent un autre type de source, de nature sonore [30]. Nous allons faire un court survol pour distinguer les deux types de systèmes et se concentrer par la suite sur les systèmes passifs.

1.3.1 Système actif

Un système actif de formation d'image va utiliser une source contrôlée pour illuminer les objets d'une scène et recueillir l'information réflective des objets. Il fait la réception de la même information qu'il émet comme l'illustre la figure 1.1. Par exemple, un sonar émet des ultrasons : l'information recueillie sera le retour des ultrasons frappant les objets de la scène par le récepteur du système. La source émise peut être un laser, un rayon X, champ magnétique, etc. Le système actif n'a pas toujours besoin d'une lumière ambiante pour être fonctionnel.

Ce type de système est très pratique pour aller chercher de l'information par rapport à des caractéristiques précises comme de l'information sur la structure de

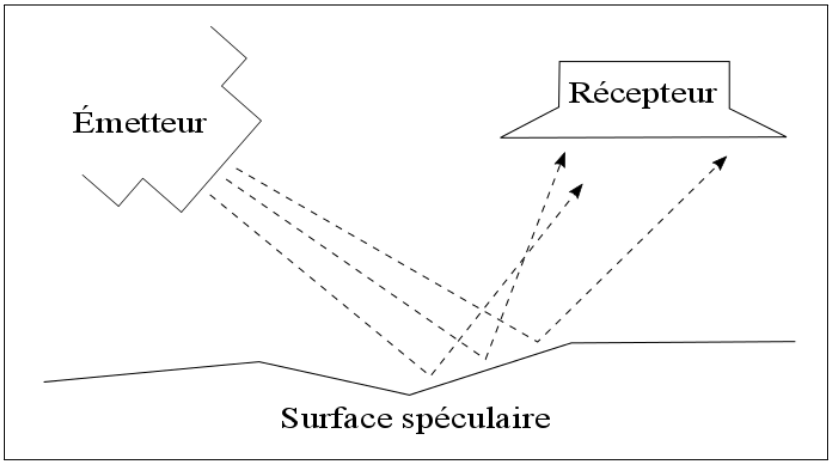


figure 1.1 – Système actif

la matière avec les rayons X. C'est donc un système très utilisé dans le domaine de la médecine [12] avec les images rayon X et aussi avec l'imagerie par résonance magnétique (IRM) qui elle peut contenir de l'information sur les tissus du corps humain [19]. L'information géométrique par rapport à la scène, comme nous amène principalement un système passif est ici mise de côté.

1.3.2 Système passif

Un système passif (figure 1.2) quant à lui, n'envoie aucune énergie à partir d'un émetteur, mais utilise l'énergie réflective qu'elle reçoit à partir de la scène. Cette énergie peut provenir d'une ou de plusieurs sources éclairantes telles qu'une lumière ambiante comme celle du soleil ou encore une source telle qu'utilisée par la tomographie par émission de positons, une technique d'imagerie médicale pratiquée par les spécialistes en médecine nucléaire.

L'information recueillie dépend des longueurs d'onde que l'on cherche à acquérir, car le capteur est sensible à des ondes de la lumière dont les longueurs sont incluses dans un intervalle précis : infrarouge (IR), ultraviolet (UV), lumière visible ou autre. Le système passif est généralement un système utilisant une transformation géométrique pour permettre la manipulation de l'information comme pour la visualisation

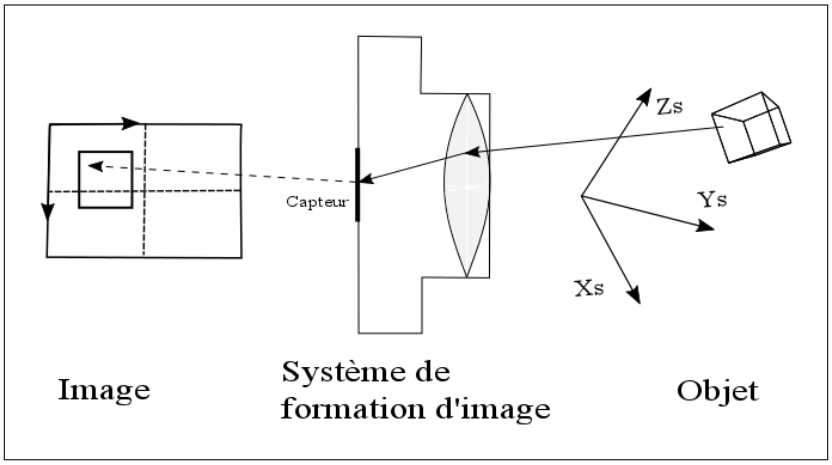


figure 1.2 – Système passif

ou le traitement d'image.

Dans la suite du document, nous allons faire l'étude de la formation d'image pour un système passif et optique uniquement. C'est-à-dire pour un système avec lentille et pour de la lumière visible qui est un rayonnement électromagnétique. Nous allons donc toujours prendre en considération que notre système de formation d'image n'émet aucune énergie. Nous allons faire la description en suivant le cheminement d'un rayon lumineux à travers tout le système de formation en partant de la source lumineuse jusqu'au capteur, en passant par la scène et la transformation en pixel pour une image. On va donc commencer par définir la lumière provenant d'une source lumineuse et les modèles de réflectance de la scène pour expliquer la provenance des rayons lumineux captés par un système de formation d'image passif.

1.4 Source lumineuse

Une source lumineuse est un objet qui va produire la lumière qu'il émet. La lumière est un phénomène physique et elle peut être définie comme étant un transport d'énergie constituée d'ondes électromagnétiques. Elle est indispensable à la vision humaine ; notre vision capte ce qu'on appelle la lumière visible, figure 1.3, qui est un ensemble d'ondes électromagnétiques dont les longueurs d'onde dans le vide sont comprises entre 380 nanomètres (violet) et 780 nanomètres (rouge) [41, 28].

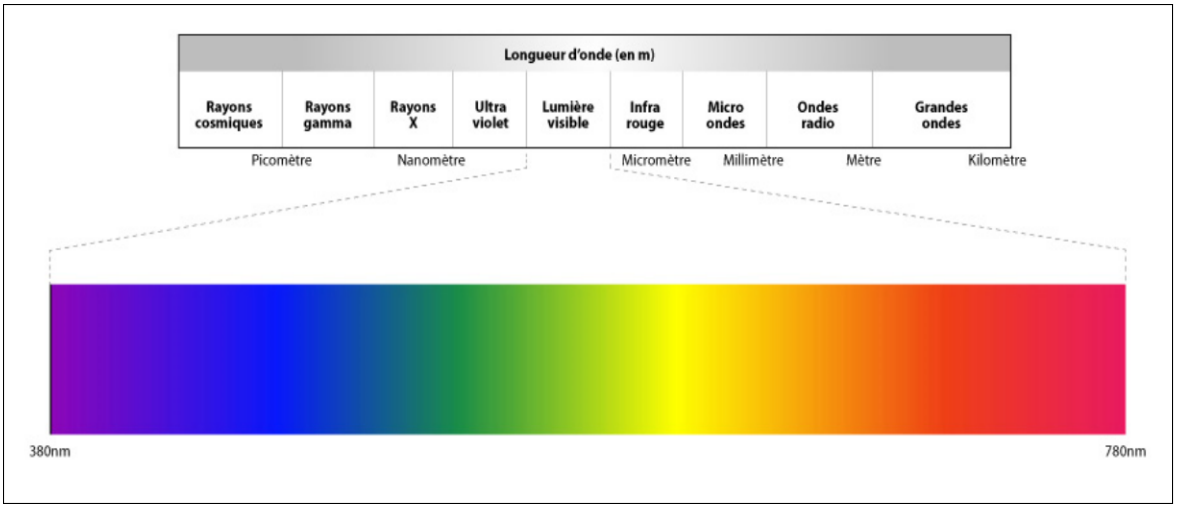


figure 1.3 – Spectre lumineux, image provenant de [36]

Pour l'humain, chaque oeil possède trois cônes qui sont des photorécepteurs transformant le signal électromagnétique de la lumière en signal nerveux permettant la sensation de couleur. Ainsi, la couleur perçue dépend des propriétés de réflectance que possèdent la scène et les objets de celle-ci. La lumière provenant d'une source qui arrive sur un objet de la scène, vois ses rayons réfléchir selon un spectre lumineux qui peut être décrit par la puissance de la distribution spectrale (PDS). Dans la scène la PDS d'un objet est une combinaison des propriétés réflectance de l'objet et de l'illumination de la scène. Chaque source lumineuse possède une PDS propre à elle-même qui est associée à la théorie des corps noirs.

1.4. Source lumineuse

Un corps noir désigne un objet dont le spectre électromagnétique ne dépend que de sa température et dont le rayonnement électromagnétique se doit d'être en équilibre thermique ; pour chaque fréquence, la radiation quittant la source est déterminée par la température d'équilibre et ne dépend pas de sa forme, de son matériau ou de la structure du corps. Cette relation est exprimée par la loi de Planck, qui dicte que le spectre émis par un corps a une intensité maximale selon une température. On peut voir avec la figure 1.4, l'intensité maximale, λ_{max}, selon le spectre visible de la température en kilojoule (K).

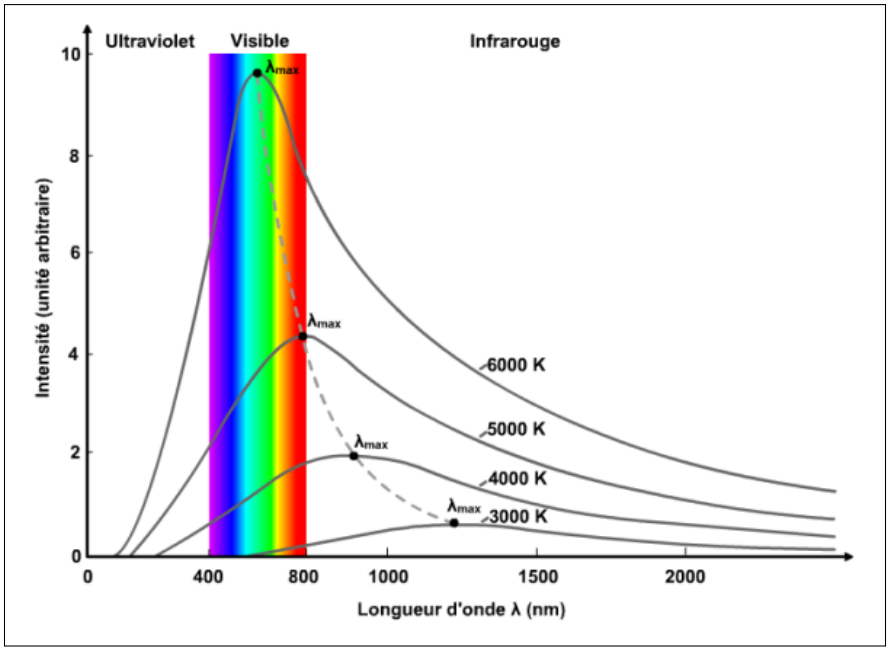


figure 1.4 – Graphique représentant l'évolution de la densité spectrale d'énergie émise par un corps noir en fonction de la longueur d'onde pour plusieurs températures

Dans cette section, nous avons parcouru le fonctionnement d'une source lumineuse en établissant un lien direct entre la couleur, la température et la longueur d'onde d'un corps. Nous allons donc, dans la section suivante, introduire une source lumineuse dans une scène et voir l'effet de la radiance spectrale sur les objets et les surfaces d'une scène par les modèles de réflectances.

1.5 Introduction à la notion de réflectance par la scène

Dans une scène, chaque source lumineuse émet des rayons lumineux dans différentes directions et chaque rayon peut être caractérisé par sa position, sa direction et sa radiance pour une longueur d'onde λ. Cet ensemble de rayons émis par la source peut aussi être appelé radiance spectrale. La lumière provenant des différentes sources lumineuses va atteindre à travers la scène les surfaces des objets rencontrés et interagir avec elles. Une surface possède alors des caractéristiques suivant la composition de ces matériaux qui peut produire trois phénomènes lorsqu'elle est atteinte par un rayon lumineux. Soit l'absorption, la réfraction et la réflexion qui sont représentées respectivement par les rayons de couleur bleu, rouge et vert dans la figure 1.5.

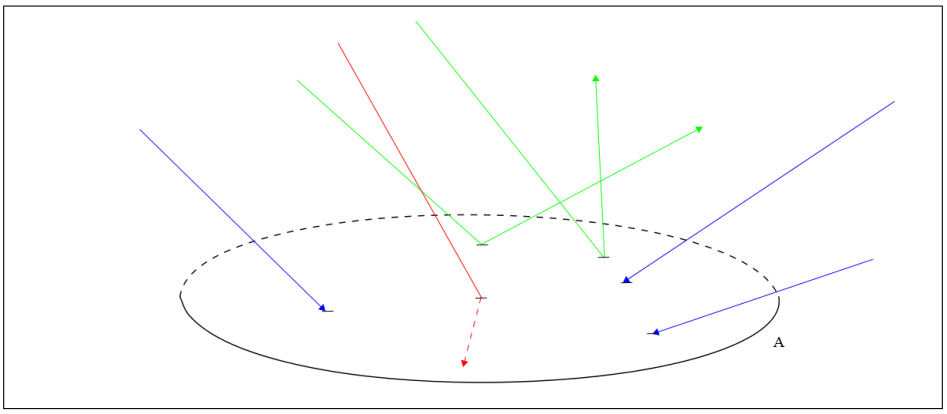


figure 1.5 – Interactions lumière/surface

L'absorption transforme l'énergie en une autre, comme en énergie thermique par exemple. La réfraction est un changement de direction qui survient à l'inverse de la normale à la surface et elle suit la loi de Snell-Descartes [40], selon les indices de réfraction n_1 et n_2 des matériaux et les angles incident du rayon par rapport à la normale à la surface i_1 et i_2 (figure 1.6) :

$$n_1 sin(i_1) = n_2 sin(i_2) \tag{1.1}$$

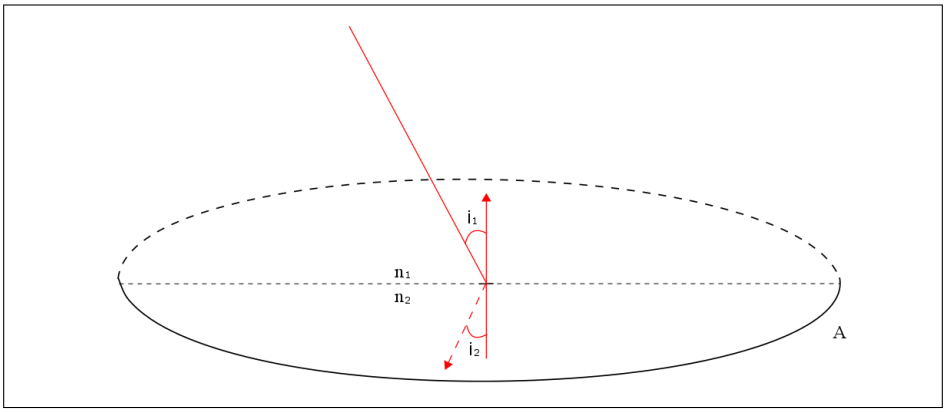


figure 1.6 – Réfraction sur une surface

puis, il y a la réflexion qui est pour nous la plus importante parce que comme pour le système visuel humain, un système de formation d'image optique va capter la réflexion spéculaire des rayons qui sont réfléchis sur la surface. Il y a deux modèles principaux de réflexion. Tout d'abord, il y a le modèle de réflexion diffuse ou de diffusion lambertienne :

$$R = Ek_d cos(i) \tag{1.2}$$

pour lequel k_d, le facteur de réflectance diffuse du matériau. Dans se modèle, un rayon réfléchi à la surface donne tout un hémisphère de directions possibles avec une certaine intensité égale pour tous les rayons [5]. Ensuite, il y a la réflexion spéculaire qui désigne la réflexion de la lumière dans une direction particulière selon un angle d'incidence. L'angle égal à l'angle d'émission du rayon arrivant à la surface sera maximisé. Les intensités perçues des rayons réfléchis différents de l'angle d'incidence dans l'hémisphère diminueront selon une loi posée par un des deux modèles suivants : le modèle de Phong [4] ou bien, le modèle de Blinn [4].

Dans le modèle de Phong l'intensité des rayons lumineux diminue selon la loi suivante : $cos(\theta)^n$. Avec θ qui représente l'angle entre $\overrightarrow{V}$, le vecteur directionnel du rayon réfléchi et celui incident sur la surface (figure 1.7). Les caractéristiques de rugosité de la surface comme pour la réfraction sont représentées par n. Plus n est grand, plus le lobe spéculaire diminue, il est donc plus local sur la surface. Plus n est

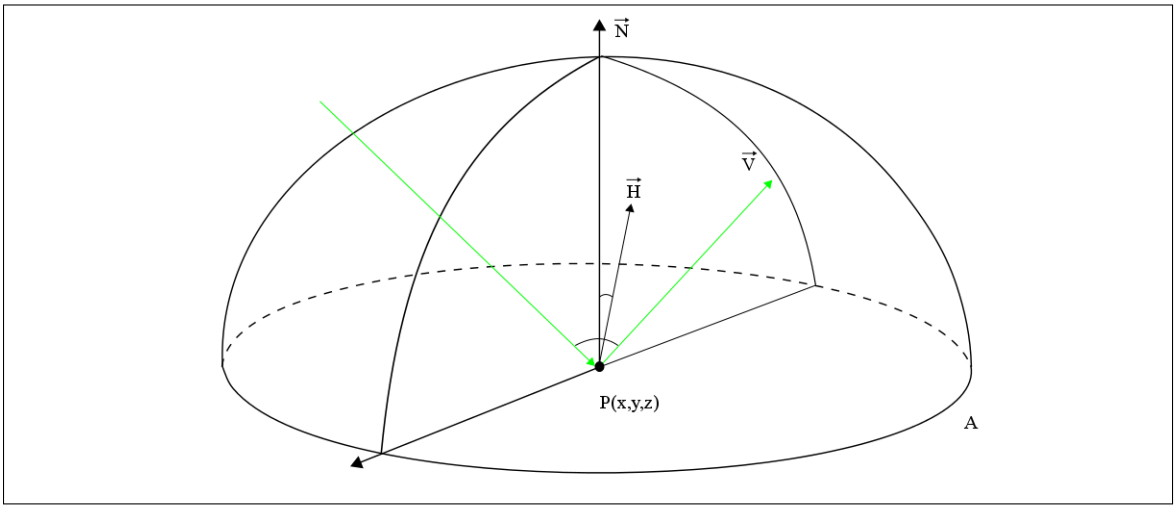


figure 1.7 – Modèle de réflexion spéculaire (Phong et Blinn)

petit, plus il est étendu sur la surface de l'objet. La seule différence entre les modèles de Phong et de Blinn est l'angle utilisé. Blinn utilise l'angle entre le vecteur $\overrightarrow{H}$ et la normale $\overrightarrow{N}$, avec $\overrightarrow{H}$ le vecteur bissecteur de l'angle θ. La radiance émise par chacun des modèles suivant la direction de $\overrightarrow{V}$:

$$R = Ek_s cos(\theta)^n \tag{1.3}$$

avec E, l'irradiance reçue par la source lumineuse et k_s le facteur de réflectance de la composante spéculaire. Le choix du modèle influence le calcul de la radiance que nous allons utiliser lors du calcul de la radiométrie pour déterminer l'intensité lumineuse dans le plan image. La luminosité est donc reliée à deux concepts, nous allons les définir avant de les introduire dans le processus de formation radiométrique de l'image dans la section 1.8.

Ces concepts résument les notions de luminosité par rapport à la scène. Avant de déterminer l'intensité lumineuse d'une image, nous devons d'abord présenter les notions de formation géométrique dans le système de formation en intégrant l'optique pour répondre à la première question posée par K.P. Horn. Par la suite, à l'aide de la radiométrie nous pourrons déterminer l'intensité lumineuse pour une image en appliquant les notions de radiance et d'irradiance.

1.6 Formation géométrique de l'image

Rappelons-nous la première question posée par Horn : qu'est-ce qui détermine l'emplacement d'un point d'un objet sur l'image ? On va donc trouver la correspondance de manière géométrique d'un point de la scène à celui d'un point dans l'image.

1.6.1 Repère de la scène et repère caméra

Tout d'abord, nous voulons avoir la possibilité de travailler dans le repère caméra pour effectuer des transformations géométriques qui vont nous permettre de transposer la scène dans le plan image, il faut donc faire un changement de repère de la scène vers celui de la caméra. Il y a alors présence de deux repères orthonormés ; un repère contenant le monde réel, le repère de la scène et un contenant le plan image, le repère de la caméra. Les deux repères possèdent leurs propres systèmes de coordonnées et un changement de repère entre celui de la scène et celui de la caméra peut se faire à l'aide des paramètres extrinsèques et intrinsèques de la caméra.

Posons $P_s = (x_s, y_s, z_s)$, un point dans le repère de la scène qui lui est défini par les axes $\overrightarrow{X_s}$, $\overrightarrow{Y_s}$ et $\overrightarrow{Z_s}$. Un point $P = (x, y, z)$ est lui décrit dans le repère de la caméra avec comme axes $\overrightarrow{X}$, $\overrightarrow{Y}$ et $\overrightarrow{Z}$. Le repère caméra est centré sur l'ouverture de la lentille de la caméra selon un axe qui est perpendiculaire au plan image, l'axe optique, qui est un axe passant par le centre de l'ouverture de la caméra en direction de notre scène. La figure 1.8 illustre tout ceci.

Ce changement de repère est possible pour tous points du repère monde en appliquant une rotation R et une translation géométrique T. Sous forme matricielle on pose :

$$P = RP_s + T \tag{1.4}$$

Une fois le changement de repère fait, il est possible de travailler dans le repère caméra pour trouver le correspondant d'un point dans le plan image. Par exemple, nous avons le vecteur $T = [t_x, t_y, t_z]^t$ qui représente les composantes de la translation

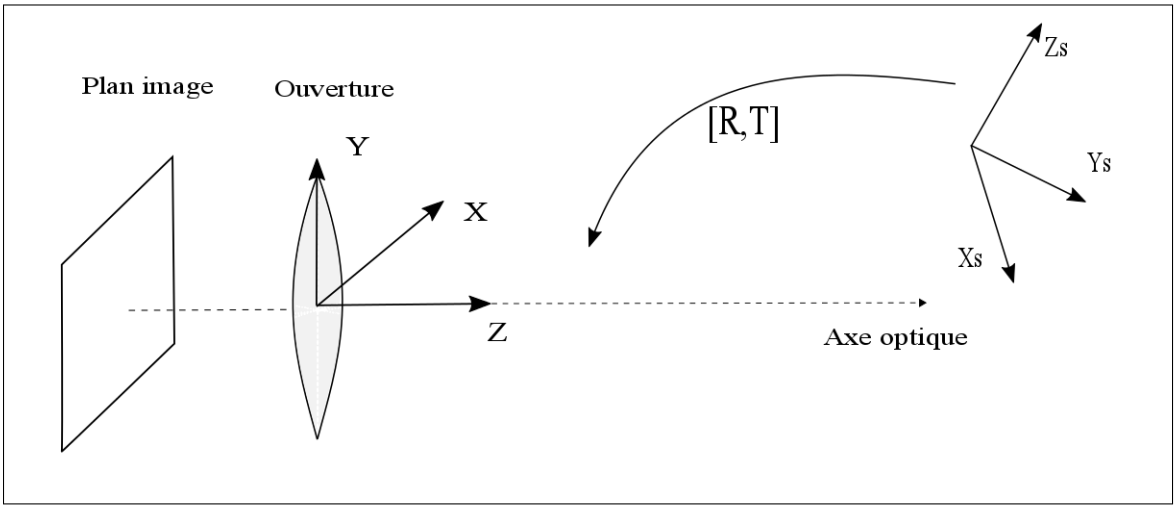


figure 1.8 – Représentation des deux repères orthonormés

liant la scène au centre de la lentille de la caméra. Il est également possible d'intégrer ces transformations sous une seule matrice globale en utilisant les coordonnées homogènes qui ont l'avantage d'être adaptées à la géométrie projective [11]. C'est-èa-dire que si deux ensembles de coordonnées sont proportionnels, ils dénotent le même point d'espace projectif. Pour cela, on doit ajouter une quatrième coordonnée généralement égale à un pour tous les points de la scène et de la caméra. On peut donc poser :

$$M = \begin{pmatrix} R & T \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} R_{1,1} & R_{1,2} & R_{1,3} & t_x \\ R_{2,1} & R_{2,2} & R_{2,3} & t_y \\ R_{3,1} & R_{3,2} & R_{3,3} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

ou M est la matrice globale de rotation et de translation représentant les paramètres extrinsèques de la caméra. Voyons maintenant comment faire les transformations géométriques qui vont nous permettre de transposer la scène dans le plan image.

1.6.2 Modèle de caméra à sténopé

Lorsque nous avons les coordonnées de la scène selon le repère caméra nous pouvons trouver la correspondance de manière géométrique d'un point de la scène à celui d'un point dans l'image avec l'aide de la projection perspective. Il existe un modèle théorique pour un cas de projection de perspective idéale qui est le modèle de caméra à sténopé. C'est à dire, sans lentille. C'est un modèle utilisé pour le calibrage géométrique des caméras et c'est en nous basant sur ce modèle que nous allons illustrer la projection perspective.

Une caméra à sténopé (figure 1.9) contient un plan parallèle situé en face du plan image par rapport à la scène. Sur ce plan, à la position de l'axe optique, il se trouve une petite ouverture, un sténopé, de la taille d'un point. La caractéristique de cette caméra est qu'un seul rayon provenant de chaque point de la scène peut traverser cette ouverture, les traits pointillés de la figure 1.9. Nous avons alors un rayon pour chaque point qui sera projeté sur notre plan image. Puisque notre axe optique consiste à être le centre de nos deux plans, si un rayon part d'une position supérieure à l'axe optique, il ne pourra traverser l'ouverture qu'en étant projeté à une position inférieure à l'axe. L'image sera donc inversée sur notre plan image par rapport à celle de la scène.

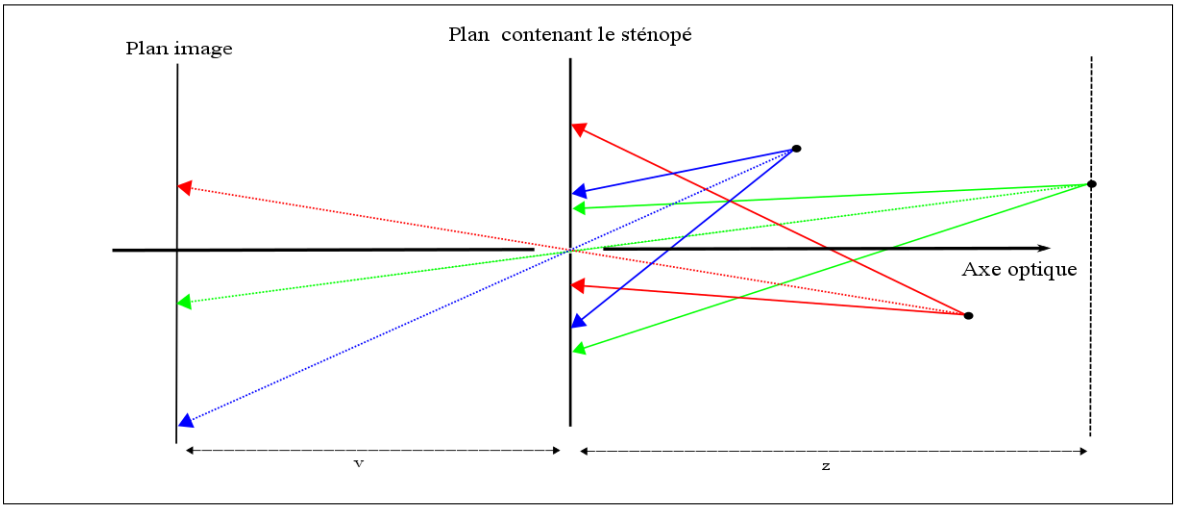


figure 1.9 – Modèle de caméra à sténopé

Posons le point $P(x, y, z)$, un point du repère caméra, qui est projeté sur un plan 2D en $P'(x', y')$. La coordonnée z' devient égale à 1. La projection se calcule en se basant sur le rapport entre d, la distance entre l'origine et le plan 2D, et la distance z qui correspond à la coordonnée dans le repère 3D. On peut alors poser le système suivant :

$$\begin{aligned} \frac{x'}{x} &= \frac{d}{z} \\ \frac{y'}{y} &= \frac{d}{z} \\ z &= d \end{aligned} \tag{1.5}$$

qui devient,

$$\begin{aligned} x' &= x\frac{d}{z} \\ y' &= y\frac{d}{z} \\ z &= d \end{aligned} \tag{1.6}$$

avec $\frac{d}{z}$ appelé terme de changement d'échelle.

Regardons encore une fois la projection perspective mais, cette fois-ci, le plan image est derrière notre origine de projection. On a v, la distance entre le plan image et le plan contenant le sténopé, qui nous permet de poser tel quel le système vu au préalable :

$$\begin{aligned} x' &= \frac{-v}{z}x \\ y' &= \frac{-v}{z}y \\ z &= d. \end{aligned} \tag{1.7}$$

Comme le démontre le modèle à sténopé, figure 1.9, l'image sur le plan est inversée selon l'axe optique. Ce même phénomène se produit avec le système visuel humain lorsque la lumière atteint la rétine. Heureusement, notre cerveau arrive à faire l'inversion pour la visualisation du monde. Présentons maintenant la lentille et ces caractéristiques dans un modèle de formation d'image.

1.7 Formation optique de l'image

Une image est un plan en deux dimensions contenant l'information lumineuse d'une scène et des objets. Pour comprendre comment acquérir l'intensité lumineuse d'un point de l'image à partir d'un point de la scène à travers un faisceau lumineux, nous allons parcourir les principes reliés à l'optique d'un système de formation d'images. Avec l'utilisation d'un système de formation d'image optique, un lien existe entre différents repères et un plan : repère de la scène, repère de la caméra et le plan image. Nous allons dans cette section voir le cheminent complet d'un faisceau lumineux avec les caractéristiques que cela implique pour chaque repère et chaque élément du système de formation d'image concerné.

1.7.1 Modèle de caméra à lentille mince

Un modèle de caméra à sténopé demeure théorique et est très peu utilisé puisque c'est un modèle qui laisse passer peu de lumière à cause de la taille de l'ouverture. La projection perspective sur ce type de modèle demeure un cas de projection idéal et les rayons passant nécessairement par le centre de l'optique ne sont pas déviés. En augmentant l'ouverture, on augmente la quantité de rayons lumineux pour chaque point de la scène, ce qui permet d'obtenir une plus grande quantité d'énergie lumineuse qui est nécessaire pour calculer l'intensité d'un point dans l'image.

Cependant, en augmentant tout simplement l'ouverture, les rayons ne convergent pas tous sur le même point du plan image, ce qui va créer des zones de floues dans notre plan image. Pour combler ce problème, il y a l'intégration de lentille, dont l'ouverture est considérablement plus grande et permet en principe de faire converger tous les rayons d'un point de la scène sur le même point du plan image par réfraction et ainsi diminuer le flou, phénomène que nous verrons en détail dans le chapitre 2.

Un modèle de caméra par lentille mince (figure 1.10) applique donc une projection perspective pour tous les rayons captés par la lentille. On appelle donc modèle à lentille mince un modèle dont nous négligeons l'épaisseur de la lentille ; c'est un modèle simplifié de lentille.

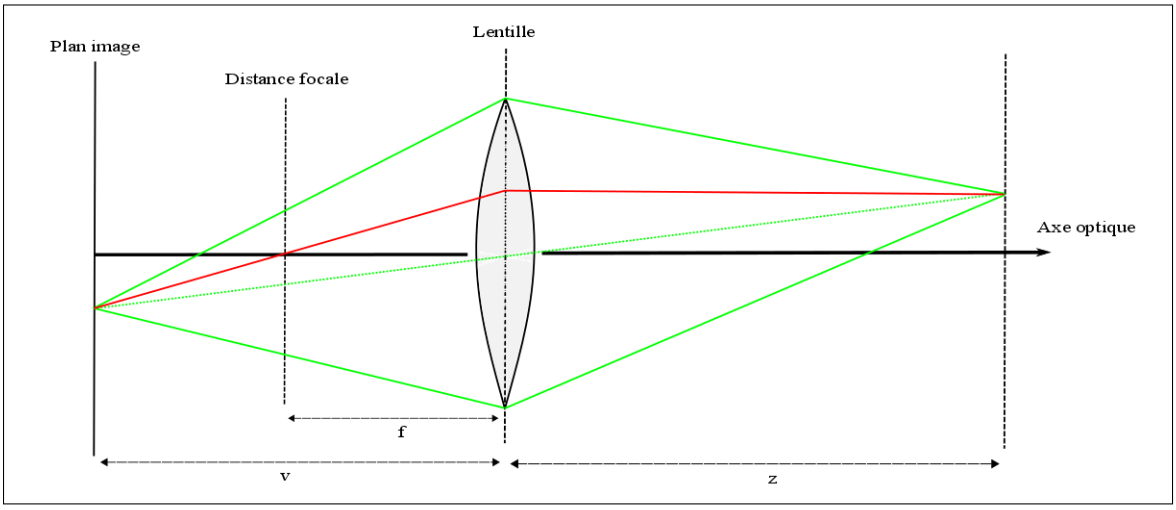


figure 1.10 – Modèle de caméra à lentille mince

Ce type de modèle possède deux propriétés intéressantes :

1. Tous les rayons passant par le centre de la lentille ne sont pas déviés.
2. Tous les rayons parallèles à l'axe optique arrivant à la lentille sont déviés par réfraction et passent ainsi par un point situé sur l'axe optique appelé foyer ou point focal (figure 1.10 - rayon rouge).

À partir de ces propriétés, on établie ce qu'on appelle la loi des lentilles minces, qui permet de vérifier à quelle profondeur un point de la scène sera projeté en un seul point sur le plan image :

$$\frac{1}{f} = \frac{1}{v} + \frac{1}{u} \tag{1.8}$$

avec comme paramètres : f la distance entre le point focal et la lentille, v la distance entre le plan image et la lentille et z la distance entre la lentille et un point de la scène. Tout point se trouvant à une distance u qui vérifie l'équation de lentille mince verra ses faisceaux lumineux se concentrer en un seul point sur le plan image. Un

point à une distance différente convergera, soit après ou avant le plan image selon la position du point de la scène par rapport à la distance focale et formera une zone de floue. Ce phénomène est illustré avec la figure 1.11 par les points $P2$ et $P3$.

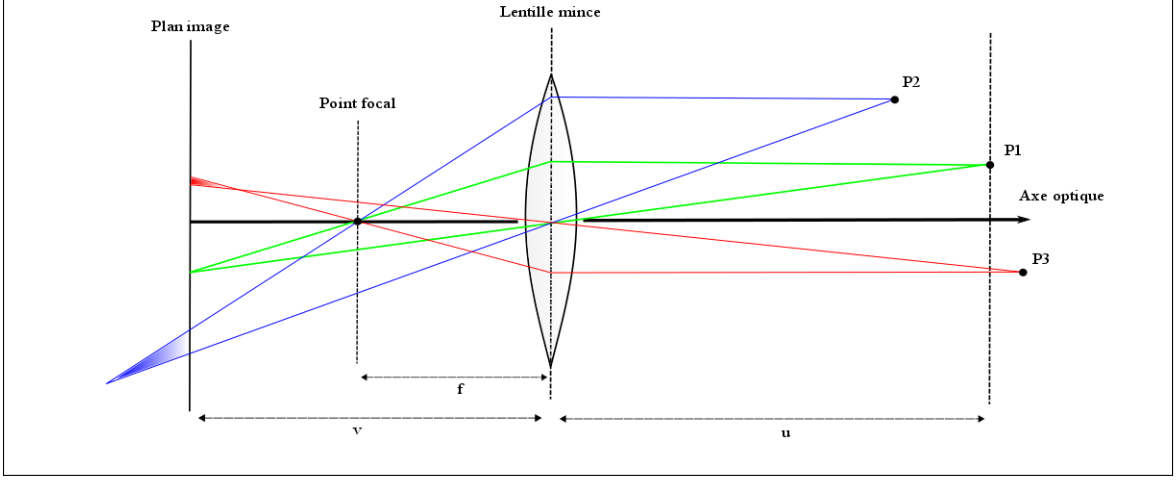


figure 1.11 – Modèle de caméra à lentille mince avec zones de flou

Nous avons répondu à la première question posée par Horn, qu'est-ce qui détermine l'emplacement d'un point d'un objet dans une image ? L'intégration de la lentille forme cependant une inversion du plan image. Nous verrons comment avoir une représentation juste de la scène après avoir fini le parcourt d'un faisceau lumineux dans le système de formation d'image. Il reste maintenant à déterminer la luminosité d'un point dans l'image.

1.8 Formation radiométrique de l'image

Maintenant se pose la question de la luminosité. Comment est-ce que l'on peut déterminer l'intensité à un point particulier de la scène dans l'image ? On va donc utiliser la radiométrie pour répondre à cette question. Par définition, la radiométrie étudie la mesure des rayonnements. Elle intègre la radiance (L) et l'irradiance (E) dans le processus de formation d'image pour calculer la puissance des rayonnements électromagnétiques sur les longueurs d'onde de la lumière visible. Dans une image la

luminosité est donc reliée au flux d'énergie qui traverse la lentille et qui est ensuite projeté sur le plan image. L'intensité de l'image finale sera déterminée par le capteur que nous verrons dans la section 1.9.

1. **La radiance (L)**

Dans la scène, la radiance est un flux d'énergie émis par une surface. C'est une puissance par unité émise (watts) d'une surface sous unité d'angle solide (stéradian) pour une surface : watts par mètre carré par stéradian ($Wm^{-2}sr^{-1}$). Plusieurs points très rapprochés d'un objet peuvent avoir une luminosité différente dépendamment de la manière dont ils sont illuminés par la source et comment la lumière est réfléchie selon les caractéristiques et la forme de la surface où ils se trouvent. Par la radiométrie nous verrons comment gérer cela pour une surface telle qu'un pixel.

Plusieurs points très rapprochés d'un objet peuvent avoir une radiance différente dépendamment de la manière dont ils sont illuminés par la source et comment la lumière est réfléchie selon les caractéristiques et la forme de la surface où ils se trouvent.

2. **L'irradiance (E)**

L'irradiance est le flux d'énergie incident sur une surface. C'est un ratio d'énergie par unité de surface, Wm^{-2} pour watts par mètre carré :

$$E(\lambda) = \frac{dP(\lambda)}{dA} \tag{1.9}$$

avec P la puissance de la lumière atteignant la surface A.

Un système optique reçoit les radiances spectrales provenant de la réflexion des différentes sources lumineuses sur une surface. L'irradiance qui arrive sur un point du plan image correspond à la projection géométrique du rayon traversant la lentille en son centre, puisqu'un rayon passant par le centre de la lentille n'est pas dévié. On établit donc un lien entre la radiance d'un point de la scène et l'irradiance de son point dans le plan image comme l'illustre la figure 1.12.

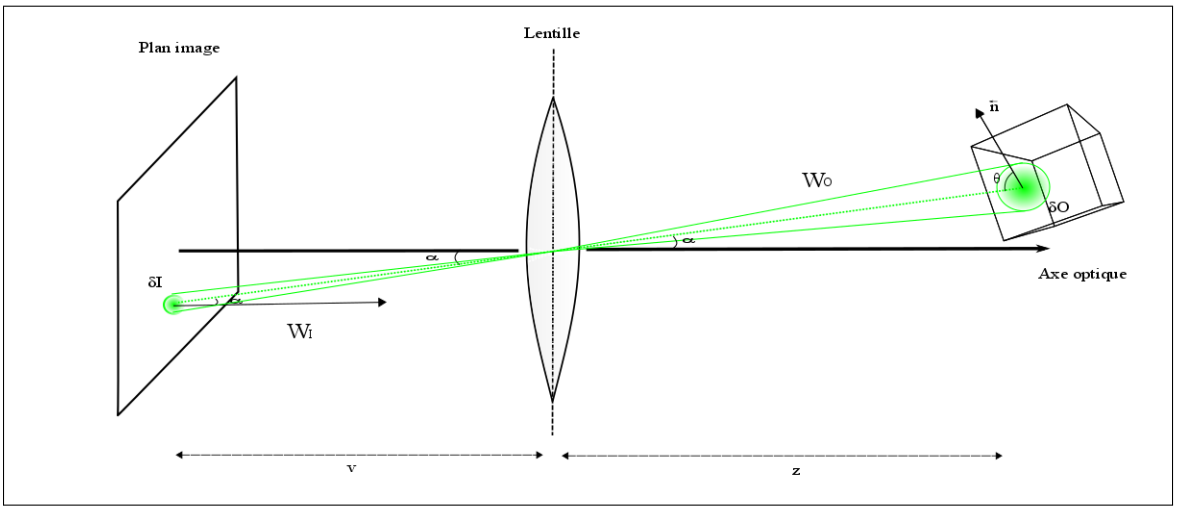


figure 1.12 – Modèle radiométrique avec angle solide

Cependant, pour un rayon lumineux, la radiance est une énergie sur une surface selon un angle solide. Pour atteindre seulement un point sur le plan image, nous allons considérer un angle solide qui devient infiniment petit, pour qu'il en soit de même avec la surface du plan image. Sinon, l'intensité lumineuse d'un point va affecter l'intensité de ces voisins dans le plan image lors de la quantification de l'irradiance. On doit donc s'imaginer que la radiance propagée par un rayon lumineux se rapproche d'un point et que la lentille est traversée par le rayon déterminé par le centre de la surface de l'angle solide. Si on se réfère encore à la figure 1.12, l'angle solide du cône des rayons arrivant à la surface de l'objet W_O est égal à celui du cône des rayons arrivant à la surface de l'image W_I. Ce qui permet de poser l'égalité suivante :

$$W_I = W_O \tag{1.10}$$

$$\frac{\delta I cos(\alpha)}{(\frac{v}{cos(\alpha)})^2} = \frac{\delta O cos(\theta)}{(\frac{z}{cos(\alpha)})^2} \tag{1.11}$$

avec $\delta I cos(\alpha)$ et $\delta O cos(\theta)$ les surfaces à partir du centre de la lentille pour l'image et l'objet. Les angles α et θ correspondent à la normale de chaque surface et $(\frac{z/v}{cos(\alpha)})^2$ est la distance entre le point dans la scène et la lentille. Cette égalité servira plus

tard pour calculer l'irradiance au plan image pour δI. Par la suite, on doit calculer la quantité de lumière émise par la surface de l'objet δO qui sera absorbée par la lentille.

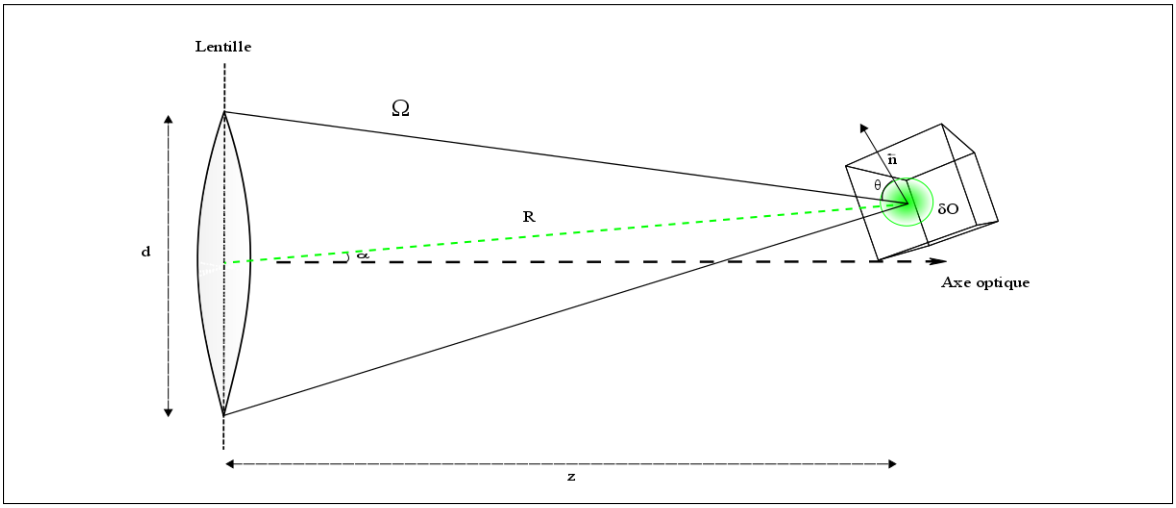


figure 1.13 – Quantité de lumière absorbée par la lentille

De la même manière, on calcule l'angle solide qui va atteindre la lentille :

$$\Omega = \frac{(\frac{\pi}{4}d^2)cos(\alpha)}{R} = \frac{(\frac{\pi}{4}d^2)cos(\alpha)}{(\frac{z}{cos(\alpha)})^2} \tag{1.12}$$

$$\Omega = \frac{\pi}{4}(\frac{d}{z})^2 cos^3(\alpha) \tag{1.13}$$

avec $\frac{\pi}{4}d^2$ la surface de la lentille.

On calcule ensuite la puissance de la lumière partant de la surface et passant par la lentille. La puissance se traduit comme étant la radiance multipliée par la quantité de lumière arrivant sur la lentille multipliée par la surface visible de l'objet par rapport au centre de la lentille :

$$\delta P = L\Omega\delta O cos(\theta). \tag{1.14}$$

Finalement, l'irradiance absorbée pour la surface δI du plan image est traduite ainsi :

$$E = \frac{\delta P}{\delta I} = \frac{L\Omega\delta O cos(\theta)}{\delta I} \tag{1.15}$$

et on remarque que l'on retrouve le rapport $\frac{\delta O}{\delta I}$ de départ, que l'on va remplacer dans l'équation :

$$E = L\frac{\pi}{4}(\frac{d}{z})^2 cos^4(\alpha) \tag{1.16}$$

ou E est l'irradiance concentré dans l'image. C'est cette énergie qui va déterminer l'intensité atteignant le plan image pour cette surface de l'image. L'irradiance sur le plan image est ensuite traduite en signal électrique à l'aide d'un capteur qui fait l'accumulation des longueurs d'onde, avant de former l'image réelle que nous pouvons visualiser. Nous allons donc poursuivre avec la fonctionnalité et les différents types de capteurs existants.

1.9 Traitement physique de l'image (capteur)

Lors de son passage à travers le système de formation d'image, les faisceaux lumineux passent dans l'optique de l'appareil pour atteindre un capteur. Le capteur reçoit l'irradiance de la scène et doit la quantifier en valeur entière pour former l'image numérique telle que nous la connaissons [33, 28]. Sur le marché il existe différents types de capteurs. Les plus communs sont les capteurs CCD (*Charge-coupled device*) et CMOS (*Complementary metal-oxide-semiconductor*). Le CMOS a l'avantage d'avoir une vitesse de lecture d'image plus rapide que le CCD [5]. La méthode de fonctionnement est cependant la même pour les deux et c'est ce que nous allons expliquer.

Pour enregistrer une image, le capteur utilise des cellules photosensibles qui captent les photons des faisceaux lumineux. Le capteur est physiquement formé d'un tableau contenant des millions de petites cellules, dont le nombre varie selon sa résolution. Avec un appareil photographique, lorsque l'on presse sur le bouton d'obturation et que l'exposition commence, chacune des cellules fait la conversion de l'énergie lumineuse en énergie électrique selon l'accumulation des photons. Une fois le temps d'exposition fini, l'accumulation arrête pour faire une conversion de l'énergie lumineuse en énergie électrique pour ensuite faire la lecture de celle-ci. L'opération de lecture est différente dans le CCD et le CMOS et permet pour chaque cellule de faire ressortir les niveaux d'intensités, qui sont ensuite transformés en valeur entière, soit en bit.

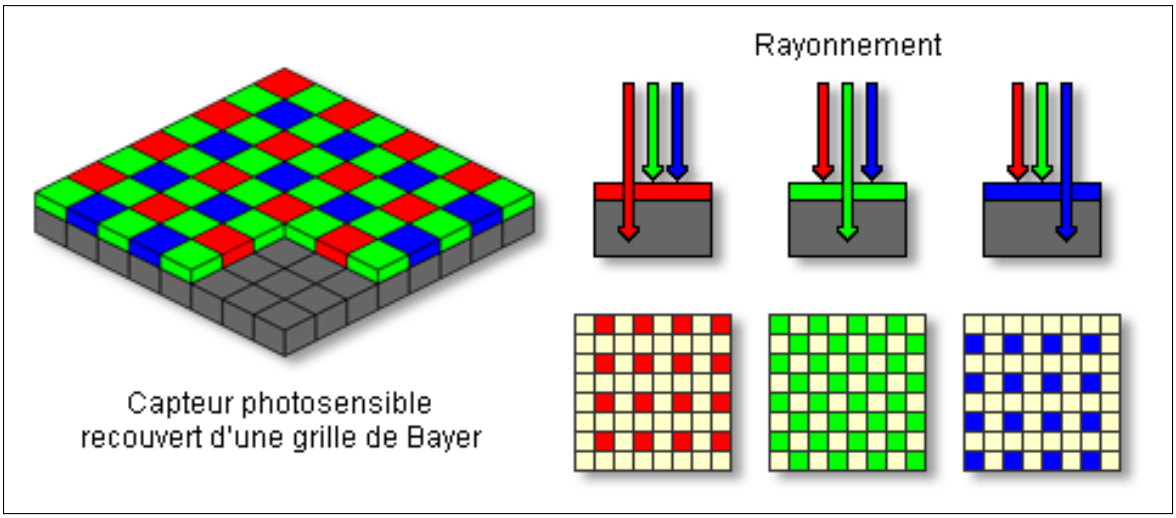


figure 1.14 – Exemple de capteur CCD

Par contre, le capteur est incapable de distinguer la longueur d'onde des photons et en faire la séparation en couleurs. C'est pourquoi un filtre doit être placé à l'entrée de chacune des cellules pour permettre à une bande couleur d'être capturée par les cellules. Ce filtre de couleurs est appelé grille de Bayer ou matrice de filtres couleur CFA (*Color filter array*). C'est un filtre qui distingue les trois couleurs primaires sensibles au spectre visible de la lumière. Chaque couleur primaire ne couvre cependant pas la même fraction de la surface du capteur. Comme on peut voir sur la figure 1.14, le vert représente les deux tiers de la surface. C'est inspiré par l'oeil humain qui est plus sensible à la lumière verte comparativement aux étendues de longueurs d'ondes primaires [31]. Bien que cette configuration soit la plus commune, il existe plusieurs configurations des filtres de couleurs [6].

Puisque seulement une bande de couleur est captée par cellule photosensible, les filtres Bayer utilisent l'interpolation pour mesurer la couleur finale dans les points où elle n'a pas été mesurée. Différentes techniques de reconstruction et d'interpolation des couleurs existent selon les types de capteurs, la plus fréquente est l'interpolation par bloc. C'est-à-dire de voir chaque bloc de deux par deux comme étant une seule cellule. Chaque cercle noir de la figure 1.15 représente un pixel et chacun des carrés

de la grille de Bayer touchés par un de ses cercles, représentent l'information qui sera interpolée pour mesurer la couleur finale du pixel.

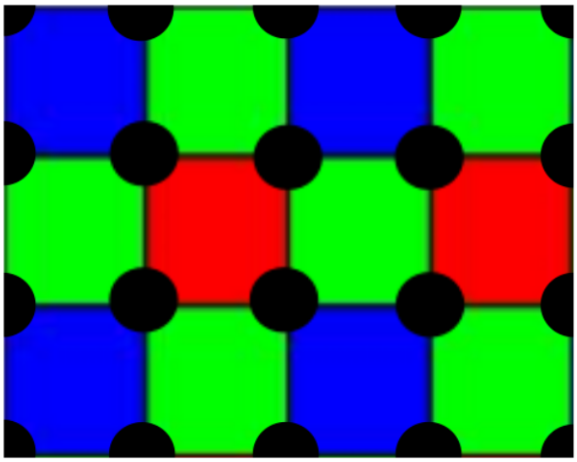


figure 1.15 – Interpolation par bloc

Cependant, une méthode d'interpolation a été proposée dans [14] pour contrer les lacunes de l'interpolation par bloc qui peut former de l'aliasing et autres artéfacts autour des contours de l'image qui sont visuellement perceptibles. Elle intègre un modèle d'interpolation spectrale inspiré par la théorie des centres de masses en physique. Cette méthode implique de faire une détection de contours avant d'effectuer l'interpolation. De cette façon, l'interpolation est faite le long des contours et non de façon uniforme, ce qui réduit la présence d'artéfacts tels que l'aliasing. Une gestion des bords doit être faite, mais généralement le calcul est tout simplement omis. Ce qui diminue la résolution de l'image comparativement au capteur mais ceci est négligeable lorsque l'on parle de millions de pixels. Maintenant que nous avons notre image numérique, nous nous retrouvons avec un système de formation d'image complet avec lequel nous voulons travailler.

1.10 Conclusion

Dans ce chapitre, nous avons parcouru les différentes étapes de la formation d'image en partant de la scène jusqu'à l'image finale produite par le capteur que nous pouvons visualiser. Nous avons également introduit que le monde qui nous entoure, varie selon l'observation et l'analyse que l'on en fait. En utilisant un système de formation optique, nous captons l'intensité de la scène par la réflexion (diffusion ou spéculaire) des rayons lumineux sur les surfaces et les objets. Plusieurs caractéristiques physiques sont rattachées à la réflexion des rayons selon la nature des composantes des objets et des surfaces de la scène. Un système optique permet d'introduire la notion du flou par la formation géométrique de l'image et la projection perspective qui pour chacun des points de la scène se verra transposé dans l'image. Pour un point de la scène, la projection se fait vers un pixel du capteur. Nous allons poursuivre ce mémoire avec le chapitre 2 qui va présenter les concepts et les terminologies sur le flou en expliquant en détails les conséquences que peuvent avoir les cercles de confusion.

Chapitre 2

Concepts et terminologies sur le flou

2.1 Introduction

Dans ce chapitre, nous allons introduire la notion de flou dans une image. Le flou qui peut être vu comme un phénomène observable, peut également s'exprimer sous la forme d'un paramètre. Noté σ [22, 44, 15, 42], il est relié à plusieurs paramètres intrinsèques de la caméra et il peut s'avérer primordial selon le type d'application souhaitée. Il est relié à la distance focale f, à l'ouverture N, au paramètre k, un coefficient de proportionnalité, ainsi qu'à la distance du plan image v et à la distance du point imagé u. Causé essentiellement par la lentille, le flou informe sur la profondeur de la scène. Lors de la construction d'un modèle d'estimation du regard dans le chapitre 3 nous allons utiliser le flou pour approximer la profondeur d'un point de la scène.

Ainsi pour pouvoir utiliser le flou, il est important de bien le définir et de bien comprendre le phénomène qu'il représente. Nous allons donc commencer par le définir dans son ensemble avant d'en voir les caractéristiques qu'il possède à travers les cercles de confusion, la profondeur de champ et les fonctions d'étalement d'un point qui peuvent établir un lien direct entre le paramètre σ et la profondeur d'un point de la scène.

2.2 Définition du flou

Comme nous avons vu avec le chapitre 1 sur la formation d'image, un point de la scène qui ne vérifie pas l'équation de la lentille mince provoque une zone de flou sur le plan image. Les variations de la quantité de flou sont causées par la variation de profondeur de la scène. Par conséquent, l'image d'un point de la scène ne vérifiant pas l'équation de la lentille mince est un disque dont le diamètre est proportionnel au flou.

Il est à noter, qu'à cause des aberrations de la lentille, tous les points projetés de la scène forment en pratique une zone de flou et qu'aucun n'est pleinement concentré en un seul point sur le plan image. Les principales aberrations connues sont l'aberration sphérique, chromatique, l'astigmatisme, le coma, la courbure du champ et les distorsions. Nous ferons cependant abstraction de ces aberrations pour le reste du document, mais il est possible de les parcourir plus en détails dans [5, 11].

2.2.1 Flou constant

On suppose le flou constant à l'intérieur d'une fenêtre de l'image. En posant comme hypothèse l'absence de la transparence dans la scène, chaque pixel est la projection d'un point de la scène. Deux points de la scène de profondeurs différentes peuvent donc être projetés sur deux pixels voisins. Le flou associé à ces deux pixels sera alors différent.

Bien que cette hypothèse ne soit pas réaliste, nous devons la poser pour ne pas être confrontés à un système qui n'est pas invariant par translation et qui serait donc très difficile à résoudre [7].

2.3 Cercle de confusion (CdC)

On a vu qu'on se retrouve alors avec l'information géométrique d'un point de la scène issue d'une projection perspective. Théoriquement, comme mentionné, un rayon passant par le centre de la lentille n'est pas dévié pour atteindre le point image mais

tous les autres rayons atteignant la lentille le seront pour atteindre le même point image. C'est à ce moment qu'entre en jeu la distance focale. La lentille a le désavantage de concentrer la lumière provenant d'un point en un point sur le plan image seulement s'il vérifie l'équation de lentille mince 1.8. Un point ne vérifiant pas cette équation verra ses rayons lumineux être projetés formant une aire sur le plan image (telle qu'illustrée à la figure 1.11), donnant ainsi naissance à un flou. Les cercles de confusion (CdC) correspondent donc à l'aire du cercle de flou pour un point de la scène ne vérifiant pas l'équation de lentille mince.

Pour calculer le diamètre d'un CdC, on peut utiliser le théorème de Thalès, qui est un théorème de géométrie affirmant que pour un plan, une droite parallèle à un des côtés d'un triangle sectionne ce dernier en triangles semblables. Ce qui va nous permettre d'établir les rapports de longueur et de proportionnalité entre les différents triangles qui se forment dans un système de formation d'image. En utilisant ce théorème à partir de la scène, on peut transposer le diamètre de la lentille pour le diamètre du CdC recherché que l'on va nommer C.

Plus concrètement, telle qu'illustrée par la figure 2.1 on a C_2 le diamètre du CdC dans la scène. À l'aide du théorème de Thalès, $C = C_2 m$ avec m un facteur d'agrandissement, qui permet d'associer le rapport de la grandeur d'un objet de la scène à son équivalent dans l'image.

On trouve la distance qui sépare le point P du plan focal par $|\overline{z} - z|$, avec z, la distance entre la lentille et la position du focus dans la scène et $\overline{z}$ la distance entre la lentille et le point P. Ensuite, avec Thalès on peut déduire que la longueur trouvée est proportionnelle à la longueur entre P et la lentille par rapport à la base, le diamètre de la lentille d, à celui de notre triangle ayant la base C_2, la base se situant à la distance z :

$$\frac{|\overline{z} - z|}{\overline{z}} = \frac{C_2}{d} \tag{2.1}$$

$$C_2 = \frac{d}{\overline{z}} |\overline{z} - z|. \tag{2.2}$$

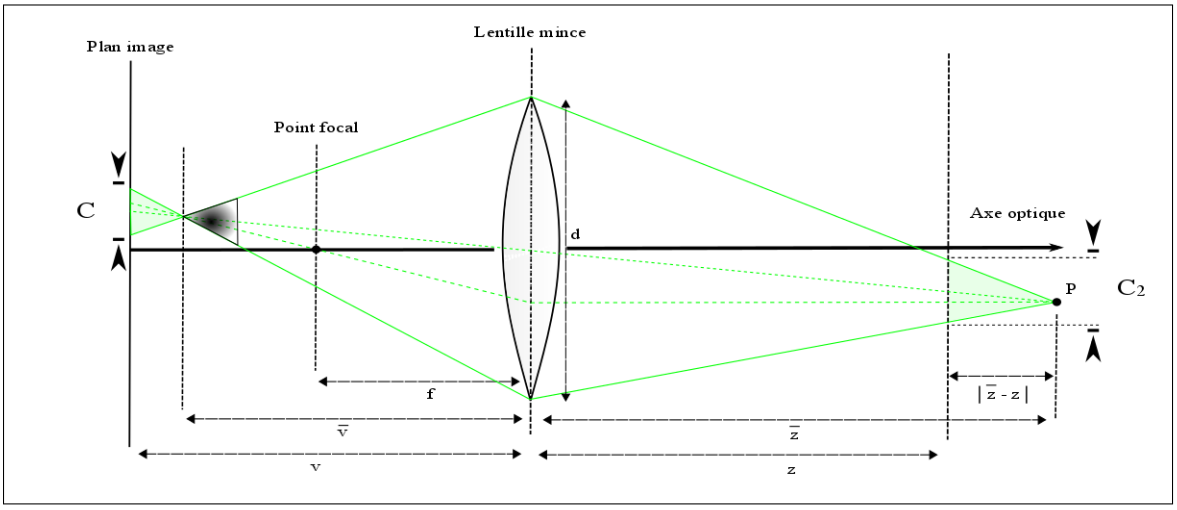


figure 2.1 – Représentation d'un cercle de confusion

Comme on peut remarquer, le triangle ayant la base C est identique au triangle noir. Il ne reste plus qu'à appliquer le facteur d'agrandissement $m = \frac{f}{f-z}$ dans le cas d'une lentille mince pour obtenir le diamètre du CdC dans l'image :

$$C = C_2 m \tag{2.3}$$

$$C = \frac{d}{\overline{z}} |\overline{z} - z| m \tag{2.4}$$

$$C = \frac{d}{\overline{z}} |\overline{z} - z| \frac{f}{f - z} \tag{2.5}$$

Les cercles de confusion sont des phénomènes importants dans la compréhension du flou dans une image. Le diamètre d'un CdC, comme on va le voir dans la section 2.4 suivante, fait partie intégrale de la profondeur de champ en servant de délimitation.

2.4 Profondeur de champ (DoF)

Pour l'oeil humain la profondeur de champ (DoF), est la distance entre le point le plus loin et celui le plus près qui permet aux objets d'apparaître acceptablement nets à notre vision. La diminution de la netteté que l'on perçoit en visualisant une scène se fait graduellement de chaque côté d'un objet qui se trouve à la distance du plan image net. Par conséquent, à l'intérieur de la profondeur de champ, le flou est imperceptible à la visualisation. En vision par ordinateur, cette zone peut être décrite à l'aide des CdC. Par exemple, si l'on est en présence d'un CdC causé par un point de la scène dont le diamètre se retrouve à l'intérieur de la résolution d'un pixel du plan image, on dira que l'on est à l'intérieur de la profondeur de champ. À l'inverse, on dit d'un point dont le diamètre d'un CdC est plus grand que la dimension d'un pixel qu'il est à l'extérieur de la DoF.

La figure 2.2 démontre le phénomène par rapport aux points $P2$ et $P3$ qui délimitent la zone de la DoF avec C, qui est de la largeur d'un pixel. On peut observer que de chaque côté du plan image, le diamètre de C diminue graduellement.

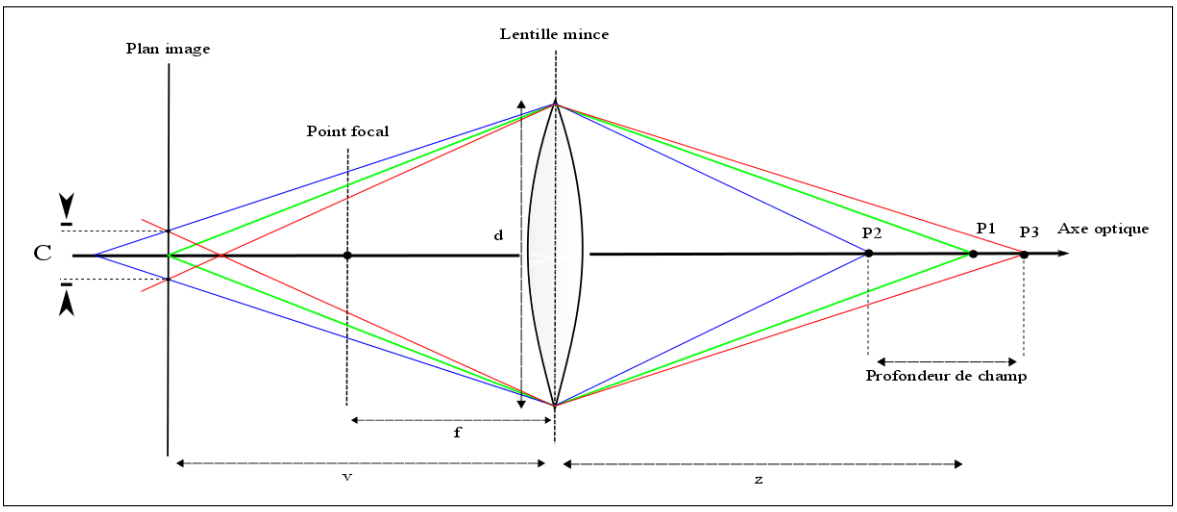


figure 2.2 – Zone de la profondeur de champ

Le calcul de la profondeur de champ est directement lié à deux paramètres intrinsèques de la caméra par rapport à la dimension fixe du capteur de la caméra : la

distance focale f et l'ouverture du diaphragme N. La figure 2.3, démontre l'impact de l'ouverture de la lentille sur la profondeur de champ pour une même distance focale. Plus l'ouverture est grande, moins la DoF est profonde. À l'inverse, moins l'ouverture est grande, plus la DoF est profonde. C représente ici le diamètre maximal d'un CdC assez petit pour être considéré négligeable.

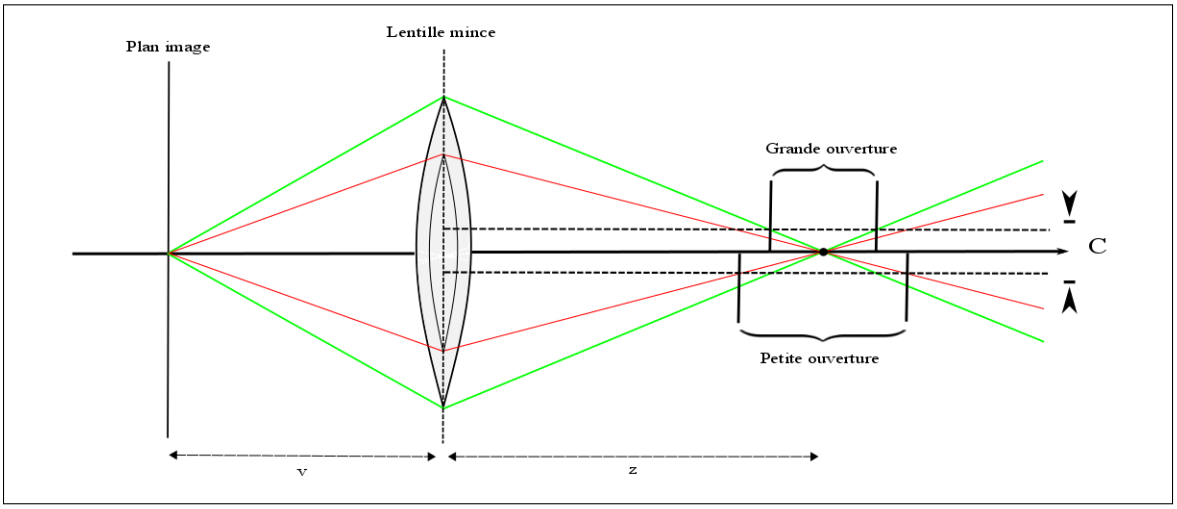


figure 2.3 – Profondeur de champ selon l'ouverture

Le même phénomène est observable par rapport à la distance focale. Pour une ouverture N et une distance v donnée, la distance entre la lentille et le plan image, plus la distance focale est courte plus la profondeur de champ est grande et l'inverse. Procédons maintenant au calcul de la profondeur de champ.

Pour le calcul de la profondeur de champ, nous avons besoin de la distance hyperfocale (H). Une distance que l'on peut définir comme étant la distance minimale entre la lentille et le plan image, pour laquelle nous avons un point suffisamment net. On pose alors :

$$H = \frac{f^2}{NC} \tag{2.6}$$

La distance H est donc entièrement dépendante du niveau de netteté que l'on considère acceptable, c'est-à-dire d'avoir le diamètre maximal d'un CdC à l'intérieur d'un pixel.

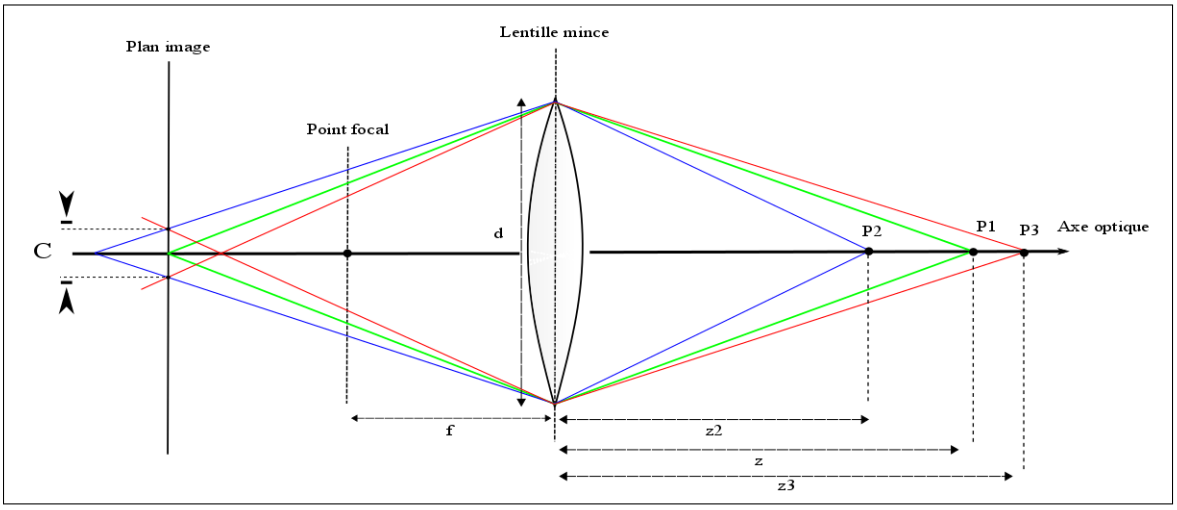


figure 2.4 – Calcul de la profondeur de champ

À partir de la distance H, on calcule z_2, la distance entre le point le plus près de la lentille qui permet un point net et z_3, la distance entre le point le plus loin qui permet un point net dans la scène :

$$z_2 = \frac{H \cdot z}{H + z} \tag{2.7}$$

$$z_3 = \frac{H \cdot z}{H - z}. \tag{2.8}$$

Tous les éléments sont réunis avec z_2 et z_3 pour calculer la zone correspondante à la profondeur de champ :

$$DoF = z_3 - z_2. \tag{2.9}$$

Finalement, pour un capteur donné, la DoF dépend seulement des paramètres f et N. Nous allons maintenant poursuivre avec les fonctions d'étalement d'un point qui viennent rejoindre encore une fois les CdC en intégrant le flou.

2.5 Fonction d'étalement d'un point (PSF)

Nous allons maintenant introduire les fonctions d'étalement d'un point (PSF). La lentille peut être modélisée à l'aide de ces fonctions. La lumière émergente de la lentille peut être vue comme la convolution entre l'image de cette lumière et une de ces fonctions [5, 8, 9].

Les rayons lumineux traversant la lentille dirvergent et interférent les uns avec avec les autres pour former de la diffraction sur le plan image. Pour une ouverture circulaire, le modèle de diffraction 2D est appelé disque d'Airy. Une PSF est donc une fonction qui caractérise la répartition de l'intensité lumineuse lorsqu'un point de la scène est projeté sur une surface. Elle est donc la fonction qui modélise la réponse impulsionnelle du système optique de la caméra. En posant une PSF pour un point $p(x, y)$, nous retrouvons l'image d'irradiance floue (I_{flou}) par une convolution avec l'image nette (I_{nette}).

$$I_{flou} = (I_{nette} * PSF)(x, y) \tag{2.10}$$

Ainsi, il est important de choisir une bonne PSF pour la simulation du flou. Dans la littérature, la modélisation d'une PSF se retrouve principalement sous deux formes [5, 8, 9] : le patron d'Airy et sous forme gaussienne. Le modèle gaussien est celui que nous allons utiliser dans le chapitre 3 et que nous allons détailler. Pour le patron d'Airy nous allons simplement énoncer la structure. Pour plus d'information on peut se référer à [5, 39].

2.5.1 Le patron d'Airy

Comme mentionné, l'effet d'une PSF est décrit par une opération de convolution. Même pour un système optique parfait, dépourvu d'aberration, la PSF n'est pas représentée par un point dans le plan image mais par un disque en raison de la diffraction causé par les longeurs d'ondes lorsqu'elles traversent la lentille. Le patron d'Airy sert donc de référence dans les systèmes optiques pour qualifier une image.

Elle offre une mesure de la meilleure focalisation possible selon une ouverture. Par exemple, si le disque d'Airy atteint les limites d'un pixel alors on aura la focalisation. Le patron d'Airy est en fait qu'une partie de la fonction PSF, qui elle, est calculée par la convolution de trois fonctions h^{dl}, h^{def} et h^{rect}. La fonction représentant le patron d'Airy et qui modélise la diffraction de la lentille est la suivante :

$$h^{dl}(x,y) = \left(2\frac{J(\frac{\pi}{N}(\sqrt{(x^2+y^2)})}{\frac{\pi}{N}(\sqrt{(x^2+y^2)}} \right)^2. \tag{2.11}$$

Mentionnons que la fonction $J(x)$ est une fonction de Bessel et que N est l'ouverture de la caméra. Le flou à l'ouverture de la caméra est représenté par la fonction h^{def}, qui est une fonction porte ronde (*Pillow Box*) :

$$h^{def}(x,y) = \begin{cases} \frac{1}{\pi(R(z))^2} & \text{si } (x^2+y^2) \leq R(z)^2 \\ 0 & \text{sinon} \end{cases} \tag{2.12}$$

ou $R(z)$ est le rayon du cercle de flou généré par un point à une distance z. La dernière fonction, h^{rec}, modélise la surface d'un pixel sur le capteur et elle est une fonction porte rectangulaire :

$$h^{rect}(x,y) = \begin{cases} \frac{1}{4W_xW_y} & \text{si } |x| \leq W_x \text{ et } |y| \leq W_y \\ 0 & \text{sinon} \end{cases} \tag{2.13}$$

avec W_x et W_y la dimension d'un pixel en millimètre. Nous n'irons pas plus loin dans la description du patron d'Airy. Poursuivons maintenant avec le modèle gaussien qui modélise la convolution des 3 fonctions ensembles.

2.5.2 Modèle gaussien

On peut voir par la représentation des courbes de niveaux, figure 2.5, la ressemblance entre le modèle gaussien et celui modélisé à l'aide du patron d'Airy.

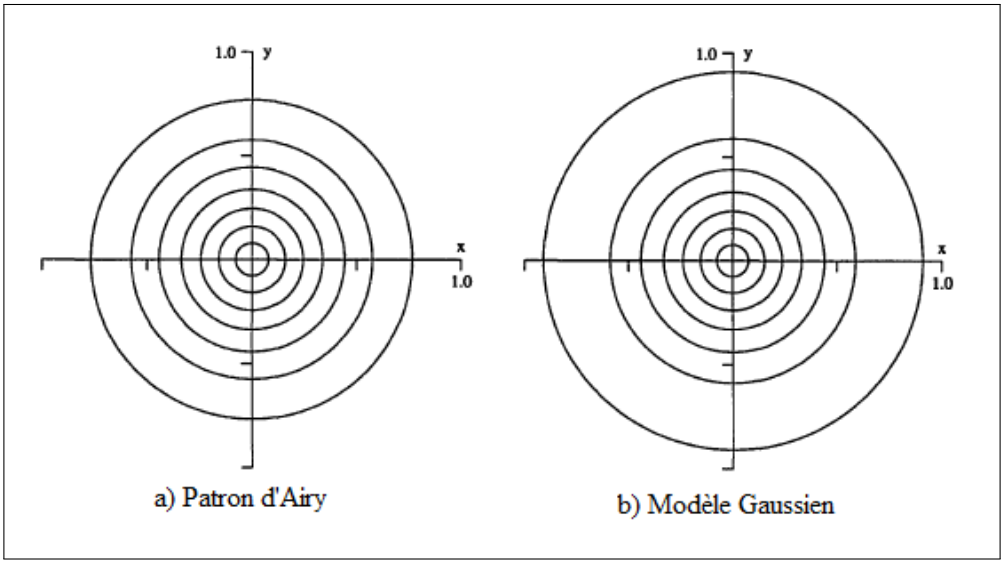


figure 2.5 – Courbes de niveaux pour le Patron d'Airy et le modèle gaussien, provenant de [39]

On a donc une équation gaussienne qui introduit le flou σ par sa variance qui dépend directement de la profondeur d'un point dans le scène :

$$g(x, y, \sigma) = \frac{1}{\pi\sigma^2} e^{-\frac{x^2+y^2}{\sigma^2}}. \tag{2.14}$$

Le flou est donc proportionnel au disque d'un CdC et avec la figure 2.6, on peut voir la relation géométrique entre un point z et σ. On peut trouver la valeur de σ par l'équation suivante qui est dérivée de [5, 45, 47].

$$\sigma = \frac{dv}{k}\left(\frac{1}{f} - \frac{1}{v} - \frac{1}{z}\right). \tag{2.15}$$

Si nous connaissons σ, on peut calculer directement la profondeur d'un point par rapport à la lentille de la caméra par l'équation 2.16 suivante qui provient de [7, 45] :

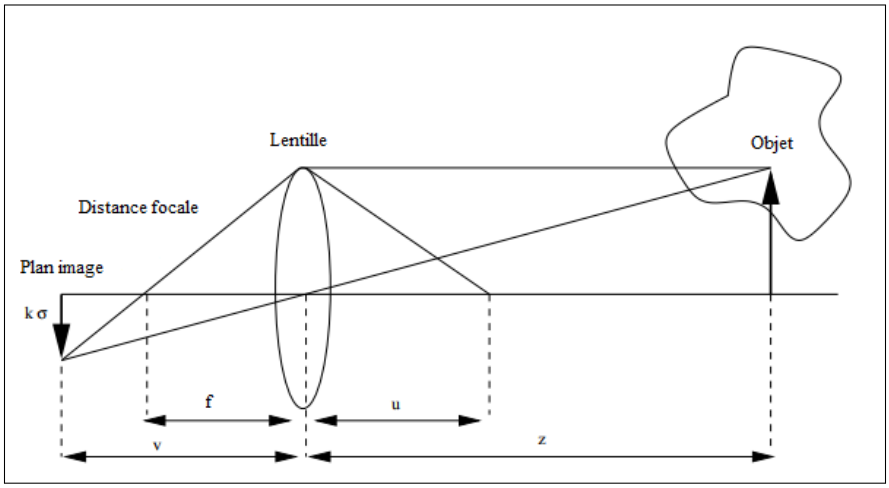


figure 2.6 – Relation géométrique entre σ et z, provenant de [45]

$$z = \begin{cases} \frac{fv}{v-f-kN\sigma} & \text{si } z > u \\ \frac{fv}{v-f+kN\sigma} & \text{si } z < u \end{cases} \tag{2.16}$$

Les paramètres N, f, v et k sont encore une fois les paramètres intrinsèques de la caméra qu'on peut aller chercher à l'aide de la calibration. Avec N, qui est l'ouverture de la caméra, f, la distance focale et v, la distance entre le plan image et la lentille. Finalement, le paramètre k, est le coefficient de proportionnalité entre l'écart-type de la gaussienne et le diamètre d'un CdC d'un point de la scène. Cette équation est valide hors foyer. C'est-à-dire que pour un point z de la scène, si nous sommes à une distance dépassant la distance au foyer dans la scène u, nous utilisons la première partie de l'équation. Inversement, si nous sommes plus près de la lentille et que la distance z est plus petite que la distance u, nous utilisons la deuxième partie de l'équation.

Dans le chapitre suivant, nous allons utiliser une fonction PSF gaussienne pour implémenter un système d'estimation du regard. Nous allons estimer la quantité de flou σ par une méthode de défocalisation.

2.6 Conclusion

Ce chapitre vient compléter le chapitre 1 sur la formation d'image. En élaborant sur les concepts impliquant le flou dans une image nous touchons au système de formation en entier puisque tous les paramètres de la caméra sont impliqués. Les cercles de confusion, la profondeur de champ et les fonctions d'étalement d'un point permettent de représenter le flou comme un indice de profondeur dans l'image. Ces concepts demeurent fondamentaux pour la compréhension du processus de formation d'image. Nous avons présenté ces concepts dans le but de construire un système d'estimation du regard à l'aide du calcul de variation de la quantité de flou dans une image. Le prochain chapitre présentera donc l'estimation du regard sur un écran d'ordinateur.

Chapitre 3

Estimation du regard

L'estimation du regard est étudiée en vision par ordinateur parce que la compréhension de la façon de regarder d'un individu peut avoir des bénéfices multiples. Par exemple, pour les chercheurs dans le domaine de la psychologie on peut, par l'observation du regard, faire l'analyse des états cognitif, l'analyse du niveau d'attention et de certains comportements humains et même, de problèmes neurologiques [35, 34, 10]. Également, pour les applications interactives, on peut élargir les possibilités concernant les interfaces homme-machine. Plusieurs études ont été réalisées en marketing [38, 27] pour étudier le comportement des consommateurs en plus du domaine de l'automobile [17, 21] qui est de plus en plus actif pour contrer la fatigue au volant. L'article correspondant à ce travail est intitulé *Real time eye-gaze estimation on a computer screen*. J'ai accompli ce travail sous la supervision du Professeur Djemel Ziou dans le but de publier dans la revue *International Conference on Image Analysis and Recognition* (ICIAR).

Real time eye-gaze estimation on a computer screen

Eric Néron and Djemel Ziou

abstract

This paper describes two approaches for eye-gaze estimation through a 2D camera such the integrated webcam laptop. The approaches differ from the head position and orientation estimation step ; one requires a marker on the front of the user, while the other one is entirely free components by using the blur in the image. A variational approximation of the Bayesian multinomial logistic regression is used to implement the gaze mapping function from head and eyes features. No calibration of the external parameters of the camera is required. The validation shows the effectivness of the proposed approaches.

Keywords: Eye-gaze estimation, marker, blur, head position, Bayesian logistic regression.

3.1 Introduction

Eye-gaze estimation is an active area in computer vision and its potential applications are numerous. For example, researchers in psychology can analyse the cognitive states, the level of attention, some human behaviors, and even some neurological disorders through the gaze [35, 34, 3]. For marketing, it can be used to target recommendations of goods and services knowing what a user is watching on the computer screen [4, 5]. For safety, it could be useful for monitoring the vigilance and fatigue of vehicle drivers [6, 7].

In the last years, recent advances in video cameras eye-gaze estimation have been investigated and several approaches have been proposed like in this survey [8]. Some of them are based on active vision systems and other on passive ones [9, 12]. For example, there are approaches in which electronic devices such as cameras equipped with active-illumination systems are used [9] and others with infrared illumination

or complex and expansive systems [13, 14]. Some of these approaches deal with head motion limitation [18] and some have tricky calibration or multiple cameras [10]. Most of the proposed approaches use passive color cameras operating in visual spectrum and 3D cameras such as the Kinect for the relief estimation of the human face [10]. The problem we are dealing with is slightly different. Indeed, we are interested in the users in front of their computer and we propose to use the embedded camera in the computer screens and the screen as a scene to watch.

Given an image of a user in front of his computer taken by the embedded camera, we would like to find the position on the screen that the user is looking at. The idea is to estimate some relevant image features and then to map these features to a position on the screen. The features include the position and orientation of the head. For this, we have experienced two different models. In the first model, a marker is placed on the forehead of the user. The 3D estimation takes advantage of the shape of the marker. The second model is based on the estimation of the user face relief from the blur contained in the image. The implementation of the later methods requires the knowledge of some intrinsic parameters of the camera. The mapping is seen as a supervised classification problem implemented by using variational approximation of the bayesian logistic regression. Note that for supervised classification, the logistic regression outperforms the support vector machine [10]. The position, orientation, eyes location, and a reference point detected automatically on the visage of the user are the feature space used by the Bayesian logistic regression. The paper is organized as follows. In Section 2 a system overview is presented. In Section 3, the eye-gaze approaches is explained. Section 4 presents experimental results and the conclusion is presented in section 5.

3.2 System overview

Let us consider that a user is watching the screen of a computer equipped with an embedded camera operating in the visible spectrum. We assume that the user environment is illuminated. The objective of a gaze estimation system is to determine what the user is watching on the screen. It means that the document components

are referenced in the screen coordinate system. These components can be a text, a word, a picture, drawing, among many others. The smallest constituant that can be succesfully predicted depends on the intrinsic and extrinsic features of the camera, the user location, the illumination, and the gaze estimation algorithms [8, 18]. This issue will be discussed in the experimental section. The area we would like to predict where the user is watching on the screen is named the area of regard (AoR). The main assumptions behind the proposed approaches are : 1) The head occupies an important area in the image ; 2) The head orientation and the gaze are near.

The first assumption is realistic because the user is generally at a distance of approximately 500 mm from the screen. A distance recommended by the Canadian Centre for Occupational Health and Safety (CCOHS) to counter the eye fatigue [19]. The second is also realistic because the user watches the screen. The general approach used for the estimation is depicted in figure 3.1. The screen is logically subdivided into different target areas. The detection of gazes consists in the definition of a mapping between a set of features of the eyes and face and the areas location in the screen. In short, given a feature vector acquired at time t, the goal is to find the AoR. This mapping can be built by using a supervised classification. Indeed, let us consider a labelled data of the form $AoR = F(\text{feature vector})$, where F is the mapping function. Several parametric and non-parametric methods can be used for the estimation of F. We propose the use of a variational approximation of Bayesian logistic regression. The logistic regression is well established method which outperforms the support vector machine [10, 15]. The feature space is formed by eight gaze parameters, which are the iris displacements d_x, d_y as well as six others parameters T_x, T_y, T_z, R_x, R_y and R_z for head position and orientation. The efficiency of these features has already been demonstrated for gaze estimation [10, 18]. The estimation of head location and orientation is described in the next section.

3.3 Mapping estimation

Let us consider the data formed by (v_i, AoR_i), where the feature vector $v_i = [R_x, R_y, R_z, T_x, T_y, T_z, d_x, d_y]_i$ and AoR_i is the coordinates of the center of the i^{th} area

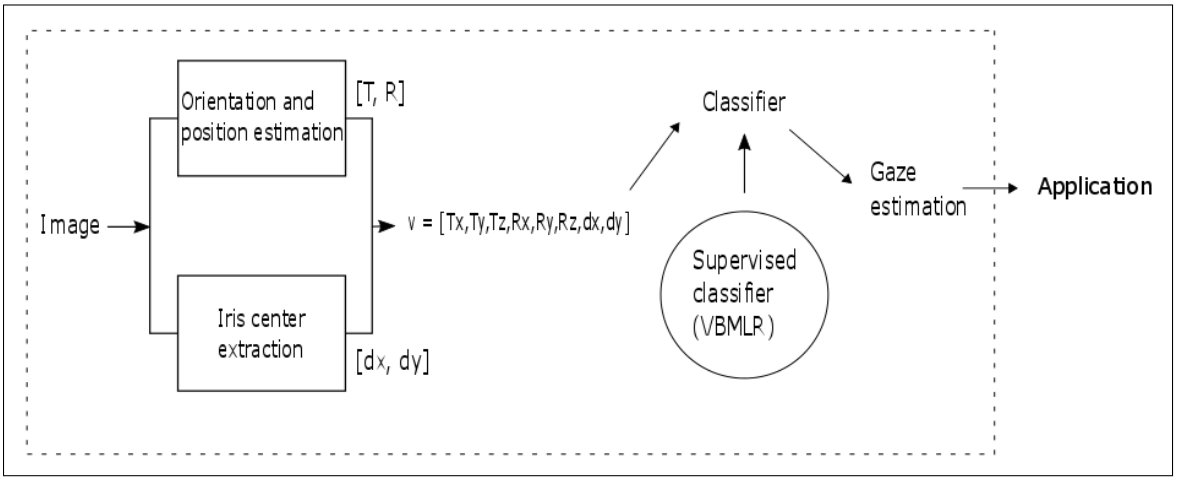


figure 3.1 – Gaze estimation system framework

of the screen. The size of an AoR is rectangular and varies according to the experiments. Let us recall that the goal is to find the mapping $v_i = F(AoR_i)$. Several mappings methods could be used such as the linear regression, neural network, and polynomial functions. We propose the use of variational approximation of Bayesian multinomial logistic regression (VBMLR). The main advantage of using a logistic regression model is that we will be able to quantify the association between each independent variables and a dependent variable Y, while taking into account the effect of the other variables integrated in the model. Making the logistic regression multinomial, will make it possible to estimate a statistical model Θ that discriminates each area from the others by the uses of a probability density function (pdf).

The mapping F is learning based. A labeled vector $v_{k,1}, \ldots, v_{K,M}$ associated with the k^{th} area is assumed to be available. The number of areas K, is already known. For a set of possible areas, represented in $\Theta = (Y_1, \ldots, Y_K)$ form, where Y_i is the set of values affecting the i^{th} area. We construct a Y_i, with respect to an amount of training data vector $v_i = [R_x, R_y, R_z, T_x, T_y, T_z, d_x, d_y]$ acquired. If we consider a complete dataset, for each vector v_i a binary variable is associated, $y_i = 1$ if it belongs to the k^{th} area, $y_i = 0$ if not. The probability of the binary variables is given by :

$$p(y_0 = 0, y_i = 1|v, \beta_i) = \frac{e^{\beta_i^T v}}{1 + \sum_{j=1}^{m-1} e^{\beta_i^T v}} \quad , i = 1, \ldots, K \tag{3.1}$$

and

$$p(y_0 = 1, y_i = 0|v, \beta_i) = 1 - p(y_0 = 0, y_i = 1|v, \beta_i) \tag{3.2}$$

where β_i is the vector of regression coefficients associated to the i^{th} area. We can classify a new feature vector v into an area by the probability associated with it and by comparing the probabilities of areas. Since we use a multinomial logistic regression, we need to estimate a model for each area. A regression is defined by :

$$Y_i = \beta_{0,i} + \beta_{1,i} x_1 + \cdots + \beta_{M,i} x_M \quad , i = 1, \ldots, K. \tag{3.3}$$

To estimate the parameter vector β_i, we used a Bayesian formulation because we want to prevent some disadvantages that can come from other coefficients estimation methods, like maximum *a posteriori* probability (MAP) and maximum likelihood estimation (MLE). The collinearity and separability of the data, the existence of a large number of zeros in the explanatory variables and even overfitting, are some examples. The implementation of a model thus passes through these $M + 1$ unknown coefficients. Let us put $p(\beta_i)$ a Gaussian prior of mean μ and covariance Σ, we need to find the parameters of the vector β_i which maximizes the posterior probability $p(\beta_i|y_0 = 0, y_i = 1)$ given by :

$$max_{\beta_i} \quad p(\beta_i|y_0 = 0, y_i = 1) \propto p(\beta_i) \sum_{x \in \Omega_i} \prod_{k \in \{0,1\}} p(y_k = k|x, \beta_i) q_k(v). \tag{3.4}$$

However, in the case of multidimensional data, the estimation of β_i fails because of the inadequacy of computational capacity of a computer when calculating the function in the eq. 4 [20]. In this case, variational approximation and Jensen equality can be used to approximate the posterior.

$$P(\beta_i|y_0 = 0, y_i = 1) \propto p(\beta_i) \prod_{k \in \{0,1\}} F(\epsilon_k) e^{(E_{q_k}(H_k) - \epsilon_k)/2 - \varphi(\epsilon_k)(E_{qk}(H_k^2) - \epsilon_k^2)} \tag{3.5}$$

where $H_k = (2k-1)\beta_k^t v_k$, E_k the expectation with respect to q_k, $\varphi(\epsilon_k) = \tanh(\frac{\epsilon_k}{2})/4\epsilon_k$, and ϵ_k a variational parameter. The approximation of the posterior above is a Gaussian with a posterior mean μ^{post} and a posterior covariance Σ^{post} given by :

$$(\Sigma_i^{post})^{-1} = (\Sigma)^{-1} + 2\sum_{k\in 0,1} \varphi(\epsilon_k) E_{q_k}(v_k v_k^t) \tag{3.6}$$

$$\mu_i^{post} = \Sigma^{post}(\Sigma^{-1}\mu + \sum_{k\in 0,1} (k-0.5) E_{q_k}(v_k)) \tag{3.7}$$

$$\epsilon_k^2 = E_{q_k}(v_k^t \Sigma^{post} v_k) + (\mu_i^{post})^t E_{q_k}(v_k^t v_k)\mu_i^{post}, k \in \{0, i\}. \tag{3.8}$$

The vector parameter β_i is the mean vector μ^{post}. For more mathematical explanation the reader is refereed to [10, 15].

3.4 Feature space

We will now deal with the estimation of the vector v. But before, we will discuss the variables used and their relevance. Each acquisition leads to a feature vector $v = [R_x, R_y, R_z, T_x, T_y, T_z, d_x, d_y]$. To find its features we need to create a reference point. A reference point is used to evaluate the spatial position of the head of the user in the 3D scene with the help of a rigid transformation based on a matrix rotation (R) and a translation vector (T). We can generate this point by using a marker placed on the user face [21].

The iris displacement d_x, d_y represents a vector from the iris center and the reference point which were measured along the horizontal and vertical axes in the image plane. The iris center and the reference point vary significantly with head position and orientation. Then, to allow head movements, the head orientation R_x, R_y, R_z and position T_x, T_y, T_z parameters are added to our mapping model. They are also calculated by using the reference point. In what follows, we will explain how these features are estimated. We begin with the iris disparities d_x and d_y, then we will explain the rotation R and translation T. Figure 2 show an example of detection we used to find

all these features.

figure 3.2 – Marker and nose tip detection example

3.4.1 Iris disparities estimation

From an image acquisition, the eye-gaze estimation needs to find the iris center. No prior knowledge or calibration is required here. In fact, after a user has oriented and positioned his head, he adjusts the eyes for a more detailed look. This adjustment passes through his iris center, which focuses precisely on the desired point in the scene. This focus can be modeled by the formation of the displacement vector $d = [d_x, d_y]$.

To calculate the displacement vector we find the difference between the iris center and the reference point. We apply Viola-Jones [17] object detection to find the eyes of the user. After eyes detection, with a crop image of one eye (see Fig. 3.3) we calculate the iris center based on references [22, 23]. Let us assume that the eyes are open. The iris detection is based on the use of its circular edge. The iris is the circle on which a maximum change of the global contrast held. More specifically, let us consider a circle of a center (x_0, y_0) and a ray r. The iris circle is the solution to the following optimization problem :

$$max_{(r,x_0,y_0)}|G_\sigma(r) * \frac{\partial}{\partial r}\oint_{r,x_0,y_0}\frac{I(x,y)}{2\pi r}ds| \tag{3.9}$$

The gaussian G_σ is used to reduce noise effects. The complete operator behaves as a circular edge detector, blurred at a scale set by σ. The absolute value is used to avoid the use of the derivative sign. For the implementation, it is more efficient to interchange the order of convolution and differentiation and to concatenate them before computing the discrete convolution [22].

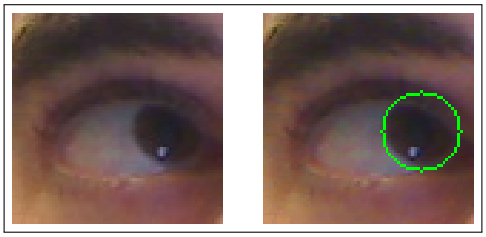
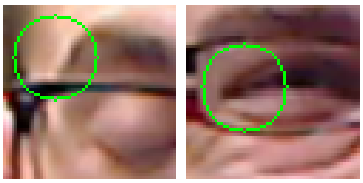


figure 3.3 – Iris detection example

Let us now discuss the assumption behind the proposed detector. The iris may not be visible in the image during blinking (Fig. 3.4a) of the eyes or when a person laughs. We find that the iris can be detected when it is partially hidden but sometimes, if the iris is too hidden and we can't detect a circle form, the iris center can be offset (see Fig. 3.4b).



(a) Blinking eye (b) Hidden iris

figure 3.4 – Iris detection for particular cases

After the calculation of the iris center coordinates $Iris_x$ and $Iris_y$, we find the iris displacement vector with the reference point :

$$\begin{aligned} d_x &= Iris_x - ref_x \\ d_y &= Iris_y - ref_y \end{aligned} \tag{3.10}$$

3.4.2 Head position and orientation using a marker

As mentioned, the head pose estimation needs a rigid transformation based on a matrix rotation (R) and a translation vector (T). We choose to use a marker based on fiducials localization technique because it efficiently solves the position and the orientation problem [25]. It is placed on the forehead of the user in range with the nose as shown in figure 3.2. The orientation and position are 3D head coordinates in the scene with respect to the center of the lens. The image plane is perpendicular to the optical axis Z.

The marker is a rhombus and all the dimensions are assumed to be known. The marker is detected in the acquired images by using the detection marker function available in the OpenCV library Aruco [24]. Let us consider $(x', y', 1)$ one of the four corners of the marker expressed in the image system reference. This point is the projection of the scene point (x, y, z) expressed in pixels (see Fig. 3.5) and the focal distance f is expressed in pixels. The projection is given by :

$$\begin{aligned} x &= \frac{z(x'-c_x)}{f} \\ y &= \frac{z(y'-c_y)}{f} \end{aligned} \tag{3.11}$$

where c_x and c_y are the coordinates of the image center. With the knowledge of z, we can express the marker in the scene system reference instead of the current pixel reference. Therefore we refer to [25, 26] to determine the z coordinate, in our case, the distance between a corner point A of the marker and the lens. The marker forms a plane and because the marker is a rhombus, we can write :

$$\overrightarrow{AB} = \overrightarrow{DC} \Leftrightarrow \begin{bmatrix} B_x & -A_x \\ B_Y & -A_y \\ B_Z & -A_z \end{bmatrix} = \begin{bmatrix} C_x & -D_x \\ C_y & -D_y \\ C_z & -D_z \end{bmatrix} \tag{3.12}$$

where A, B, C and D the four corners of the marker and $\overrightarrow{AB}$ and $\overrightarrow{DC}$ two parallel vectors. We can replace with eq. 3.11 each corner by its projection in the image plane.

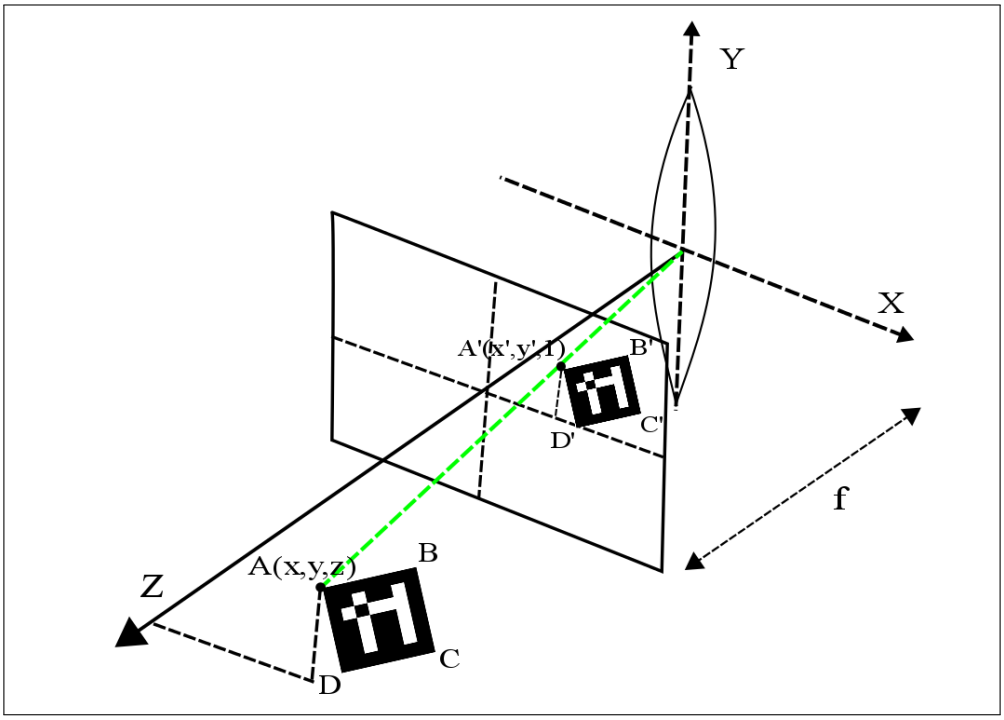


figure 3.5 – Marker coordinates system between the scene and the image plane

$$\begin{bmatrix} B'_{x'}B_z & -A'_{x'}A_z \\ B'_{y'}B_z & -A'_{y'}A_z \\ B_Z & -A_z \end{bmatrix} = \begin{bmatrix} C'_{x'}C_z & -D'_{x'}D_z \\ C'_{y'}C_z & -D'_{y'}D_z \\ C_Z & -D_z \end{bmatrix} \tag{3.13}$$

We next calculate the relative distance by setting the corner coordinate A_z to $A_z^{rel} = 1$ and renaming the coordinates B_z and C_z by respectively B_z^{rel} and C_z^{rel}. The relative distance is an arbitrary depth related to the real depth by a scaling factor. To calculate the depth of the marker from the relative depth, it is necessary to find the scale factor :

$$\begin{bmatrix} B'_{x'} & -C'_{x'} & D'_{x'} \\ B'_{y'} & -C'_{y'} & D'_{y'} \\ -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} B_z^{rel} \\ C_z^{rel} \\ D_z^{rel} \end{bmatrix} = \begin{bmatrix} A'_{x'} \\ A'_{y'} \\ 1 \end{bmatrix}. \tag{3.14}$$

After some mathematical operations (described in [25, 26]), we obtain a linear system of equations :

$$
\begin{aligned}
d &= (C'_{x'}D'_{y'} - C'_{y'}D'_{x'}) + (D'_{x'}B'_{y'} - B'_{y'}D'_{y'}) + (B'_{x'}C'_{y'} - C'_{x'}B'_{y'}) \\
B_z^{rel} &= \frac{A'_{x'}(C'_{y'}-D'_{y'})+A'_{y'}(D'_{x'}-C'_{x'})-(C'_{x'}D'_{y'}-D'_{x'}C'_{y'})}{d} \\
C_z^{rel} &= \frac{A'_{x'}(B'_{y'}-D'_{y'})+A'_{y'}(D'_{x'}-B'_{x'})-(D'_{x'}B'_{y'}-B'_{x'}D'_{y'})}{d} \\
D_z^{rel} &= \frac{A'_{x'}(B'_{y'}-C'_{y'})+A'_{y'}(C'_{x'}-B'_{x'})-(B'_{x'}C'_{y'}-C'_{x'}B'_{y'})}{d}.
\end{aligned}
\tag{3.15}
$$

Therefore the scaling factor can be expressed as a ratio. We obtain $ratio_1$ and $ratio_2$ by the following equation :

$$
\begin{aligned}
ratio_1 &= \frac{A_z}{C_z} = \frac{A_z^{rel}}{C_z^{rel}} \\
ratio_2 &= \frac{B_z}{D_z} = \frac{B_z^{rel}}{D_z^{rel}}.
\end{aligned}
\tag{3.16}
$$

With these ratios we can recover the real distance of the corner A_z. Using eq. 3.11 we have the relation between A and A', the scene point and its projection expressed in pixels. The same process is applied to corner C :

$$
A_x = \frac{A_z(A'_{x'}-c_x)}{f} \quad A_y = \frac{A_z(A'_{y'}-c_y)}{f} \tag{3.17}
$$

$$
C_x = \frac{C_z(C'_{x'}-c_x)}{f} \quad C_y = \frac{C_z(C'_{y'}-c_y)}{f} \tag{3.18}
$$

Knowing the marker dimensions, the real distance A_z of our corner A can be calculated using the ratio r_1 as follows :

$$
\begin{aligned}
f(ratio_1) &= \frac{ratio_1(C'_{x'}-c_x)-(A'_{x'}-c_x)}{f} \\
g(ratio_1) &= \frac{ratio_1(C'_{y'}-c_y)-(A'_{y'}-c_y)}{f} \\
A_z &= \frac{\|\overrightarrow{AB}\|}{\sqrt{(r_1-1)^2+(f(r_1))^2+(g(r_1))^2}}
\end{aligned}
\tag{3.19}
$$

where c_x and c_y are intrinsic parameters describing a pixel point that is at the image center. Following the same reasoning C_z is know. Likewise, B_z and D_z are known by using $ratio_2$. T is directly set by the coordinates of $A(x, y, z)$. With eq. 3.20 we have the rotation matrix expressed by r_1, r_2 and r_3. To find the orientation R we only need the normal perpendicular to the marker plane, this is what indicates the direction towards the lens, what we gives r_3.

$$r_1 = \frac{\overrightarrow{AB}}{\|\overrightarrow{AB}\|}, r_2 = \frac{\overrightarrow{AC}}{\|\overrightarrow{AC}\|}, r_3 = r_1 \wedge r_2 \tag{3.20}$$

3.4.3 Head position and orientation using the blur

We will now propose an alternative to the use of a marker in order to remove all the external elements. The main idea behind this approach is the use of the image blur. Let us recall that the blur σ is linked to the scene depth [29] by the distance z, the camera aperture N, the distance from the lens to the image v, the distance between the lens and the position of focus in the scene u, by k the proportionality coefficient between the blur circle and σ and finally, the focal distance f. In this section, all the distances are expressed in mm.

$$z = \begin{cases} \frac{fv}{v-f-kN\sigma} & \text{if } z > u \\ \frac{fv}{v-f+kN\sigma} & \text{if } z < u \end{cases} \tag{3.21}$$

It has been used for a 3D general scene reconstruction with some success [31, 34] and to approximate the depth of some elements of the scene [32, 33]. The main issue related to the use of the blur is to design a fast and accurate estimator. To fulfil this requirements, we propose the use of an estimator of edge blur [44] because at the position of the edges, it's often where we can find the discontinuity of the depth. The blur estimated on edges is propagated to the non-edge pixels. More precisely, let us consider that the model of blurred image I_b from original image I is given by :

$$I_b = I(x, y) * g(x, y, \sigma_0), \tag{3.22}$$

where g is the PSF of the camera and σ_0, the blur parameter set to an abitrary value such as 1 in our experiments. Several PSFs exist in the state of art, among them the

gaussian is most used [30, 29] :

$$g(x, y, \sigma_0) = \frac{1}{\pi\sigma_0^2} e^{-\frac{x^2+y^2}{\sigma_0^2}}. \tag{3.23}$$

Because the ledge z depends on the pixel position, then the blur σ depends on the pixel location. The estimation of the blur is carried out at each edge pixel by considering its neighbours. With the eq. 3.22, the image I_b is obtained and its contours have a more pronounced blur than the image I. Then, we calculate the ratio r between the magnitude of the gradient of the two images as follows :

$$r = \frac{\|\nabla I(x, y)\|}{\|\nabla I_b(x, y)\|} \tag{3.24}$$

$$r = \frac{\sqrt{\nabla I_x^2 + \nabla I_y^2}}{\sqrt{\nabla I_{bx}^2 + \nabla I_{by}^2}}. \tag{3.25}$$

In the case of step edge model, at the edge location, the image gradient is maximal and depends on the image contrast and the amount of blur. The ratio r is free from the contrast and depends only of the blur of the two images I and I_b [34, 35]. Therefore, the ratio is maximum at the edges, enabling us to calculate the amount of blur at this position :

$$r = \sqrt{\frac{\sigma^2 + \sigma_0^2}{\sigma^2}} \tag{3.26}$$

and

$$\sigma = \frac{1}{\sqrt{r^2 - 1}} \sigma_0. \tag{3.27}$$

Thereby we can replace the unknown σ in the formula 3.21 to find z. With the knowledge of the intrinsic parameters, the distance u can be calculated and we can use the first or the second part satisfying the condition of the formula 3.21.

Having the distance z between the nose tip and the camera lens, we will deal with the estimation of R and T. First of all, we need a reference point. By using

figure 3.6 – Nose tip and head center detection

Viola-Jones face detection [17], we can find the nose tip (blue circle inside the green box in Fig. 3.6) $N(x, y, z)$ and the head center (middle of the blue box in Fig. 3.6). Viola-Jones allows detection of the nose once the face detection is done, from which we can find the reference point $Ref(x, y, z)$ in the image plane. Eq. 3.21 allow us to find the 3D scene coordinates for these points. We assume that the user is always in front of the camera so the head center in the scene is always further than the nose tip from the lens. In this way, we add 150 mm to the nose distance to approximate the head center coordinates (see Fig. 3.7) ; $Ref_z = N_z + 150$. The nose point is used to calculate the orientation of the head by calculating the vector $\overrightarrow{R}$, the direction create between the reference point and the nose tip in the scene. The position T, is once again the reference point.

3.5 Experiments

In the experiments, each of the images acquired have dimensions of 640 x 480 pixels for width and height. The intrinsic parameters of the camera are obtained by a calibration procedure ; OpenCv calibrate camera function [27]. The calibration give to us an average re-projection error of 0.412 pixel and in the interval of 0 and 1 it's considered as a good calibration [28]. We find the focal distances $f_x = 535.127$, $f_y = 540.370$ and $cx = 297.956$, $cy = 245.628$ in pixels. The OpenCV calibration

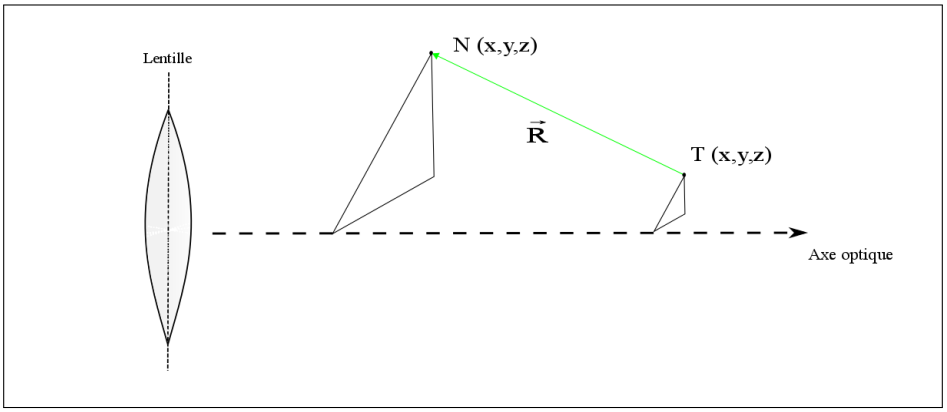


figure 3.7 – Representation of the reference point $T(x, y, z)$ and the nose tip $N(x, y, z)$ in the scene

function give to us a focal distance integrates with a scaling factor. We know the sensor size is 4.8 x 3.6 mm respectively for the width and the height. If we multiply the ratio of the focal distances and the image size by the sensor size, we find f_x and f_y in mm. We use the average of these two focal distances $f = 4.033$ mm. The distances $u = 500$ mm and $v = 4.06759$ mm. The camera aperture is $N = \frac{f}{d}$, from which we can find $k = 0.00135$. The iris detection is done with a standard deviation of the gaussian $\sigma = 0.5$. The user distance is a limit fixed to 500 mm from the camera. We used Logitech Quickcam Sphere AF camera, a standard webcam for which we have set the iris opener $d = 2$ mm.

The Pc 17 inches screen was divided into nine areas as shown in figure 3.9, each of 12.3 cm x 7.83 cm. One user participated to the training phase required for the estimation the mapping function. A user is asked to look at the center of the area on each area for a number of acquisition, the program acquisition proceeds approximately at 300 acquisitions per minute. For each acquisition, the program stores in a database the reference point and the iris center coordinates. With this information we can extract all the features. A total of 250 sample were used for training and up to 50 (# column in the experimentation tables) for tests. A blinking or different state eye, like an hidden iris whose circular shape is no longer there, can explain the difference between testing samples. The performance criteria are the number of areas accurately

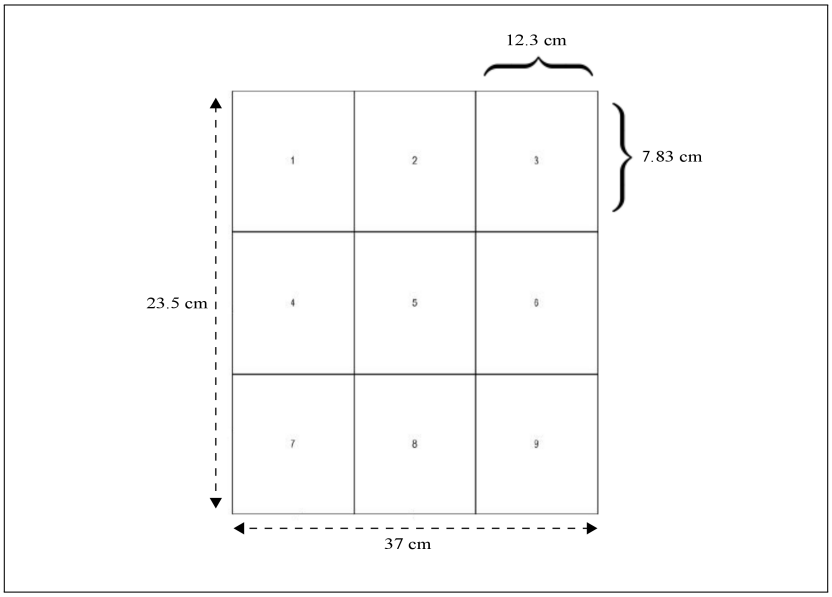


figure 3.8 – 17 inches screen subdivision

predicted as being watched by the user, the number of confusions between areas, and the computational time. Table 1 and 2 show the scores for the first experimentation. We can observe that marker method is more accurate than the blur method. The marker method is more precise because of the marker spatial positioning sensitivity. This sensitivity affects the confusion between close areas, specifically those in the same horizontal line if we compare with the blur method, as show in table 3.2 regions 2-3 and 8-9. The vector direction between the nose tip and the head center in the blur method does not offer the same flexibility of movement in the horizontal way due to the estimation of the reference point by OpenCV which is less sensitive at the head movement. In the horizontal way, by taking the middle point of the nose boxe detected (in green Fig. 9), the nose tip may be a little offset to the real tip (blue circle Fig. 9).

We repeated the same process with four different users and a 30.5 inches screen. In this configuration the size of each area was of 17.5 cm x 19 cm. We acquired 300 images per person. For each subject, we randomly extract 150 images for each region of observation. We therefore formed a training set of 4 x 150 images for each of the

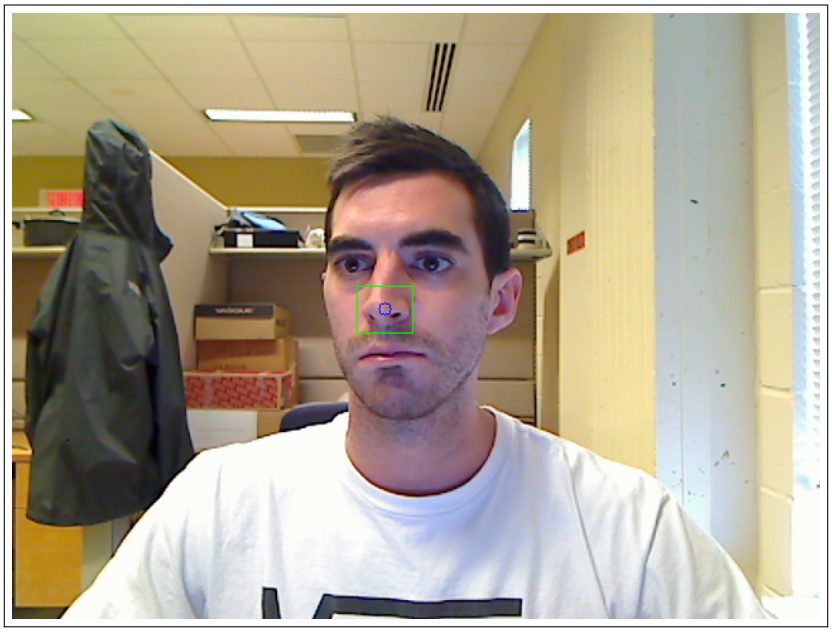


figure 3.9 – Nose detection

regions. This gives us a set of 4 different people and a total of 600 images for each region. In the remaining images for each subject, we took a sample of 75 images per region as a test set. Table 3.3 shows the scores. We can observe with respect to the blur method experiment 1 (table 3.2) a general accuracy increase in the percentage. Area enlargement brings more flexibility to the movements of the head and more precision as expected. However, the addition of several subjects can affect the pdf of certain areas ; an area can be more affected by some parameters compared to the others of the feature space, caused by the unique way each user looks. A user for example may tend to tilt more the head for the lower regions than others.

Experiment 1 : Marker method

Classification / AoR	1	2	3	4	5	6	7	8	9	%	#
1	**48**									100	48
2		**48**	1							97.9592	49
3			**50**							100	50
4		2		**37**	10					75.5102	50
5					**50**					100	49
6						**50**				100	50
7							**49**	1		98	50
8								**50**		100	50
9									**50**	100	50
Total										**98.6666**	**446**

tableau 3.1 – Training = 250, Testing samples (#) $\leq$ 50

Experiment 1 : Blur method

Classification / AoR	1	2	3	4	5	6	7	8	9	%	#
1	**50**									100	50
2		**48**	1							97.9592	49
3			**49**							100	49
4				**36**	3		8			76.5957	47
5					**31**		18			63.2653	49
6						**45**	3		1	91.8367	49
7					11	3	**36**			72	50
8								**37**	13	74	50
9								4	**46**	92	50
Total										**85.2952**	**443**

tableau 3.2 – Training = 250, Testing samples (#) $\leq$ 50

We can conclude from these experimentations that our eye-gaze system can accurately estimate the *AoR* for 9 areas on a 17 inches screen Pc for one user. Future tests have to be done with more *AoR*. Multiple users in the database decreases the overall scores compared to one subject. If we refereed to [10], we think that more subjects with an sufficient training set for each user can bring our system user invariant due to the effectiveness of the Bayesian logistic regression model.

3.6 Conclusion

In this article, we presented a eye-gaze estimation system that requires only a 2D camera such the integrated webcam compared with some existing gaze methods which use multiple cameras. Furthermore, by the estimation of the relief face from

Experiment 2 : Blur method

Classification / AoR	1	2	3	4	5	6	7	8	9	%	#
1	**27**	48								36	75
2		**68**	7							90.6667	75
3			**75**							100	75
4	3			**66**		4	2			88	75
5					**71**	4				94.6667	75
6						**75**				100	75
7				9			**66**			88	75
8								**71**	4	94.6667	75
9									**75**	100	75
Total										**88**	**675**

tableau 3.3 – Training = 600, Testing samples (#) $\leq$ 75

the blur, we remove the necessity for external calibration or components such as a marker. Experimental results showed that, despite the higher accuracy of the marker method, that blur method can reach 94 % with the experimentation 2. In the future, some new experiences will follow to improve the precision of the gaze starting with the increase number of users in the database.

Bibliographie

[1] Underwood, G. M. (2009), Cognitive Processes in Eye Guidance : Algorithms for Attention in Image Processing., Cognitive Computation 1 (1), 64-76.

[2] Geoffrey Underwood. : Eye Guidance in Reading and Scene Perception. Elsevier Science Ltd, (1998)

[3] Andrew T. Duchowski. : Eye Tracking Methodology : Theory and Practice. Springer-Verlag New York, Inc., (2007)

[4] Rik Pieters, Edward Rosbergen et Michel Wedel. : Visual Attention to Repeated Print Advertising : A Test of Scanpath Theory. Journal of Marketing Research, 36(4) :424–438, (1999)

[5] Michel Wedel, Rik Pieters. : Eye Fixations on Advertisements and Memory for Brands : A Model and Findings. Marketing Science, 19(4) :297–312 (2000)

[6] Wen-Bing Horng, Chih-Yuan Chen, Yi Chang et Chun-Hai Fan. : Driver fatigue detection based on eye tracking and dynamk, template matching. : IEEE International Conference on Networking, Sensing and Control. (2004)

[7] Qiang Ji, Xiaojie Yang. : Real-Time Eye, Gaze, and Face Pose Tracking for Monitoring Driver Vigilance. : Real-Time Imaging, 8(5) :357 – 377. (2002)

[8] D. W. Hansen et Q. Ji. : In the Eye of the Beholder : A Survey of Models for Eyes and Gaze : IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(3) :478– 500, March (2010)

[9] X. L. C. Brolly et J. B. Mulligan. : Implicit Calibration of a Remote Gaze Tracker. Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on, pages 134–134, June (2004)

[10] R Jafari, D Ziou. : Gaze estimation using Kinect/PTZ camera : Dans ROSE 2012 : Magdeburg, Germany, pages 13–18, (2012)

[11] J. G. Wang, E. Sung and Ronda Venkateswarlu, "Eye gaze estimation from a single image of one eye," Proceedings Ninth IEEE International Conference on Computer Vision, Nice, France, 2003, pp. 136-143 vol.1.

[12] Mansanet Sandin, J. ; Albiol Colomer, A. ; Paredes Palacios, R. ; Mossi García, JM. ; Albiol Colomer, AJ. (2013). Estimating Point of Regard with a Consumer Camera at a Distance. En Pattern Recognition and Image Analysis. Springer Verlag. 7887 :881-888.

[13] F. Pirri, M. Pizzoli and A. Rudi, "A general method for the point of regard estimation in 3D space," CVPR 2011, Providence, RI, 2011, pp. 921-928.

[14] Kentaro Takemura, Yuji Kohashi, Tsuyoshi Suenaga, Jun Takamatsu, and Tsukasa Ogasawara. 2010. Estimating 3D point-of-regard and visualizing gaze trajectories under natural head movements. In Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10). ACM, New York, NY, USA, 157-160.

[15] R. Ksantini, D. Ziou, B. Colin and F. Dubeau, "Weighted Pseudometric Discriminatory Power Improvement Using a Bayesian Logistic Regression Model Based on a Variational Method," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 2, pp. 253-266, Feb. 2008.

[16] Kyung-Nam Kim and R. S. Ramakrishna, "Vision-based eye-gaze tracking for human computer interface," Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on, Tokyo, 1999, pp. 324-329 vol.2.

[17] OpenCV documentation, http://docs.opencv.org/2.4/modules/objdetect/doc/cascade_classification.html

[18] R. Valenti, J. Staiano, N. Sebe, T. Gevers. Webcam-Based Visual Gaze Estimation. In International Conference on Image Analysis and Processing 2009.

[19] CCOHS. «CCOHS : Canadian Centre for Occupational Health and Safety». http ://www.cchst.ca/oshanswers/ergonomics/office/monitor_positioning.html.

[20] Djemel Ziou and Reza Jafari. 2014. Efficient steganalysis of images : learning is good for anticipation. Pattern Anal. Appl. 17, 2 (May 2014), 279-289.

[21] T. Miyake, T. Asakawa, T. Yoshida, T. Imamura and Z. Zhang, "Detection of view direction with a single camera and its application using eye gaze," 2009 35th Annual Conference of IEEE Industrial Electronics, Porto, 2009, pp. 2037-2043.

[22] J. G. Daugman, "High confidence visual recognition of persons by a test of statistical independence," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 11, pp. 1148-1161, Nov 1993. doi : 10.1109/34.244676

[23] T. A. Camus and R. Wildes, "Reliable and fast eye finding in close-up images," Object recognition supported by user interaction for service robots, 2002, pp. 389-394 vol.1.

[24] S. Garrido-Jurado, R. Muñoz-Salinas, F.J. Madrid-Cuevas, M.J. Marín-Jiménez, Automatic generation and detection of highly reliable fiducial markers under occlusion, Pattern Recognition, Volume 47, Issue 6, June 2014, Pages 2280-2292, ISSN 0031-3203

[25] Jean-Yves Didier, Fakhr-Eddine Ababsa, Malik Mallem. : Hybrid Camera Pose Estimation Combining Square Fiducials Localization Technique and Orthogonal Iteration Algorithm. : International Journal of Image and Graphics, World Scientific Publishing, 2008, 8 (1), pp.169–188.

[26] Fakhr-eddine Ababsa and Malik Mallem. 2004. Robust camera pose estimation using 2d fiducials tracking for real-time augmented reality systems. In Proceedings of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry (VRCAI '04). ACM, New York, NY, USA, 431-435.

[27] OpenCV documentation, camera calibration and 3d reconstruction. http://docs.opencv.org/2.4/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html

[28] OpenCV documentation, Camera calibration With OpenCV. https://docs.opencv.org/3.0-beta/doc/tutorials/calib3d/camera_calibration/camera_calibration.html

[29] Djemel Ziou, Francois Deschenes, Depth from Defocus Estimation in Spatial Domain, Computer Vision and Image Understanding, Volume 81, Issue 2, 2001, Pages 143-165, ISSN 1077-3142,

[30] S. Stallinga and B. Rieger, "Accuracy of the Gaussian Point Spread Function model in 2D localization microscopy," Opt. Express 18, 24461-24476 (2010).

[31] Ashutosh Saxena, Sung H. ChungAndrew Y. Ng. : 3-D Depth Reconstruction from a Single Still Image : International Journal of Computer Vision, Pattern Recognition, 76(1) :53–69, 2008.

[32] F. Deschenes, Djemel Ziou et P. Fuchs. « Improved Estimation of Defocus Blur and Spatial Shifts in Spatial Domain : A Homotopy-Based Approah ».Pattern Recognition, 36(9) :2105–2125, 2003.

[33] F. Deschenes, Djemel Ziou et P. Fuchs. « An Unified Approach for a Simultaneous and Cooperative Estimation of Defocus Blur and Spatial Shifts ». Image and Vision Computing, 22(11) :35–57, 2004.

[34] Shaojie Zhuo, Terence Sim, Defocus map estimation from a single image, Pattern Recognition, Volume 44, Issue 9, September 2011, Pages 1852-1858, ISSN 0031-3203, https ://doi.org/10.1016/j.patcog.2011.03.009.

[35] Y. Xiong and S. A. Shafer, "Depth from focusing and defocusing," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, 1993, pp. 68-73.

Chapitre 4

Résultats expérimentaux

4.1 Introduction

Dans ce chapitre, nous allons faire l'analyse des performances et des résultats du calcul de la profondeur selon nos deux méthodes et de l'application du système d'estimation du regard. En premier lieu, nous allons évaluer la profondeur z par rapport à un point de la scène pour les deux méthodes. Soit celle du marqueur et celle utilisant le flou σ. Nous ferons ensuite l'analyse des résultats pour le suivi du regard et l'analyse des performances selon les différentes étapes de l'application.

Les paramètres intrinsèques sont obtenus par l'entremise d'OpenCV et sa fonction de calibrage. Pour cette calibration, nous avons une erreur moyenne de reprojection pour un point 3D de la scène vers le plan image de 0.412 pixel. Une erreur se situant entre 0 et 1 est considérée, selon OpenCV [26], comme bonne. Il est à noter, que le modèle de formation d'image utilisé ne prend pas en compte une hypothèse de défaut de focale dans la lentille. On ne va également pas considérer un flou anisotrope, se qui serait inutile dans notre cas puisque l'on va se concentrer principalement sur le flou du visage, qui se trouve à être en avant-plan. Les paramètres d'expérimentations sont les suivants, la distance focale $f = 4.033$ mm, la distance entre la lentille et le plan image $v = 4.06759$ mm et celle entre la lentille et la position du focus dans la scène $u = 500$

mm. L'ouverture de l'iris correspond à 2 mm et l'ouverture du diaphragme $N = \frac{f}{d}$. La constante $k = 0.00135$ peut être acquise de plusieurs façon. De un, par le calcul du rayon maximal d'un CdC à l'intérieur d'un pixel. La dimension d'un pixel est dans notre cas de 0.0028 mm de hauteur et de largeur. Ou encore, en calculant le rayon de la limite de diffraction du disque d'Airy pour une longueur d'onde λ moyenne du spectre lumineux de 550 nm et une ouverture N. Dans les deux cas, la différence est minime. Chacune des images acquises dans les expérimentations ont une dimension de 640 x 480 pixels respectivement pour la largeur et la hauteur. Commençons maintenant l'analyse des résultats pour le calcul de la profondeur z avec marqueur suivi par la méthode utilisant l'estimation du flou σ.

4.2 Profondeur z avec marqueur

Nous allons premièrement présenter les résultats obtenus pour le calcul de la profondeur d'un point de la scène à l'aide d'un marqueur. Le but est ici de vérifier l'exactitude de la profondeur d'un point de la scène en utilisant la méthode du marqueur présentée dans la sous-section 3.4.2. Nous avons effectué l'acquisition de 500 images pour 18 distances différentes et calculé l'erreur qui est la différence entre la distance obtenue (X) et la distance réelle (Z). Chaque distance possède donc un échantillon de 500 images pour un total de 9000 images. Pour le calcul de la distance réelle, un marqueur est immobilisé à une distance fixe. Les distances réelles sont entre 125 mm et 550 mm. Les erreurs d'estimation sont échantillonnées par un pas de 2.5 mm quand l'erreur est plus petite que 15 mm et de 5 mm sinon. Nous avons choisi de faire un changement de pas parce que plus la distance calculée Z s'approche de la distance réelle X, plus nous voulons avoir une grande précision entre la différence des deux distances. Au-delà de 15 mm, nous avons jugé inutile que le pas d'erreur reste aussi petit l'erreur augmente avec la distance réelle. Le tableau 4.1 présente les résultats.

Comme on pouvait s'y attendre, plus la distance Z augmente, plus l'erreur augmente. Considérant que nous avons une ouverture fixe de la caméra d'à peine 2 mm, le calcul de profondeur avec l'aide d'un marqueur est relativement précis si on considère comme acceptable une erreur en bas de 10 mm et comme très bonne une erreur en bas

Erreur $(X - Z)$ / Distance réelle (mm)	[0 - 2.5[	[2.5 - 5[	[5 - 7.5[	[7.5 - 10[	[10 - 12.5[	[12.5 - 15[	[15 - 20[	[20 - 25[	[25 +
125	100								
150	100								
175	99.6			0.2	0.2				
200	97.8	2.2							
225	63.4	36.6							
250	58.2	41.8							
275	96.4	3.6							
300	22.4	41.2	25	9	0.4				
325	35.4	18.6	9.4	11.4	15.8	8.8	0.6		
350	29.8	52.4	15.8	1.8	0.2				
375	22.4	35.2	20.8	1.6					
400	20.4	21.6	27	16.4	13.4	3.6	1.2		
425	25	39.2	23	7.6	3.8	0.8	0.6		
450	23.2	19.6	12.2	13	6.4	2.6	5.6	8	9.4
475	22.6	25.4	17.8	13.4	10.2	5.2	4.4	1	
500	15.8	11.2	17.6	10.8	11.8	10.2	13.8	6.4	2.4
525	13.2	10.6	14.2	14.2	15.6	13.2	13.6	5.8	3.6
550	15.6	15.8	17	10.4	7	8.6	15.2	9.2	1.2

tableau 4.1 – Résultats en % de la classification de l'erreur pour 500 images à une distance X.

de 5 mm. Remarquons qu'à partir d'une distance de 300 mm le pourcentage d'erreur commence à s'étendre et qu'en haut de 500 mm, 40% de l'échantillon a une erreur de plus de 10 mm. Pour chaque distance passée 150 mm une variation de l'erreur est présente bien qu'il n'y a aucun mouvement du marqueur et de la caméra. Nous supposons que puisque la distance entre le marqueur et la lentille augmente, plus le marqueur devient petit dans l'image et le facteur d'échelle perd ainsi également de la précision.

Ce qui nous intéresse cependant est la stabilité et la précision de z lors du suivi du regard. Pour le test suivant, pour 3 sujets qui se retrouvent dans notre base de données, nous avons acquis 300 images et nous avons suivi l'évolution de z pour chacun. C'est-à-dire que un sujet fixe une région de l'écran et 300 images consécutives sont enregistrées par vidéo avec notre application. Un sujet est placé devant la caméra à une distance d'environ 500 mm. Cette distance peut varier d'au plus 40 mm selon le positionnement de la tête du sujet dans le repère de la scène. La posture et la stature influencent par exemple la distance puisque le sujet, lors de l'acquisition des données, est laissé à lui-même ayant comme seule indication d'être dans une position confortable. L'environnement d'acquisition sera expliqué en détails dans la section 4.4 consacrée aux expérimentations concernant le suivi du regard. Les résultats sont dans

le tableau 4.2, qui évalue pour 3 sujets la distance z pour une séquence d'image d'acquisitions.

Sujet	Médianne	Moyenne	Minimum	Maximum	Variation
1	493.1220	492.9666	477.2930	500.6770	23.3840
2	496.0580	495.2362	479.0940	500.8010	21.7070
3	481.8840	480.9250	456.9610	496.7110	39.75

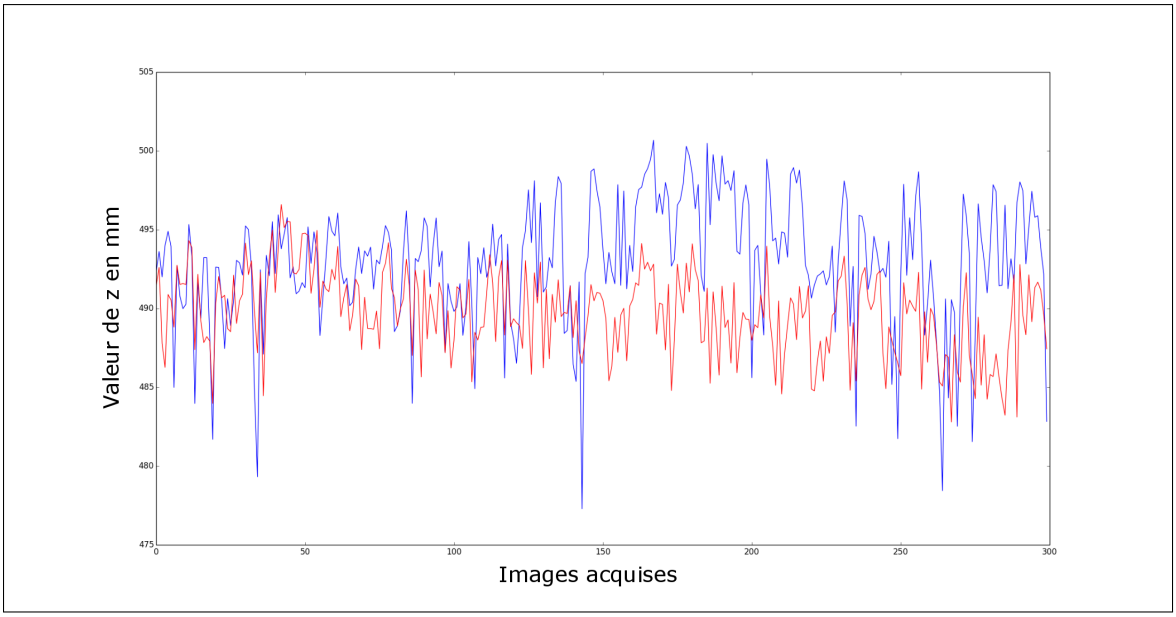
tableau 4.2 – Évaluation de z en mm pour 3 sujets à travers les images acquises

On observe que les sujets 1 et 2 oscille autour de 500 mm et semble comparable en tout point. Il y a une variabilité d'un peu plus de 20 mm entre le maximum et le minimum ce qui est acceptable si on compare avec le tableau 4.1 et la distance correspondante. Le sujet 3 laisse suggérer un mouvement du corps ou de la tête puisque la variabilité est considérablement plus grande. La moyenne plus basse est peut-être simplement causée par un positionnement initial de la tête plus près de l'écran. Nous avons tracé sous forme de graphique la représentation de z par rapport aux images acquises des 3 sujets dans la figure 4.1.

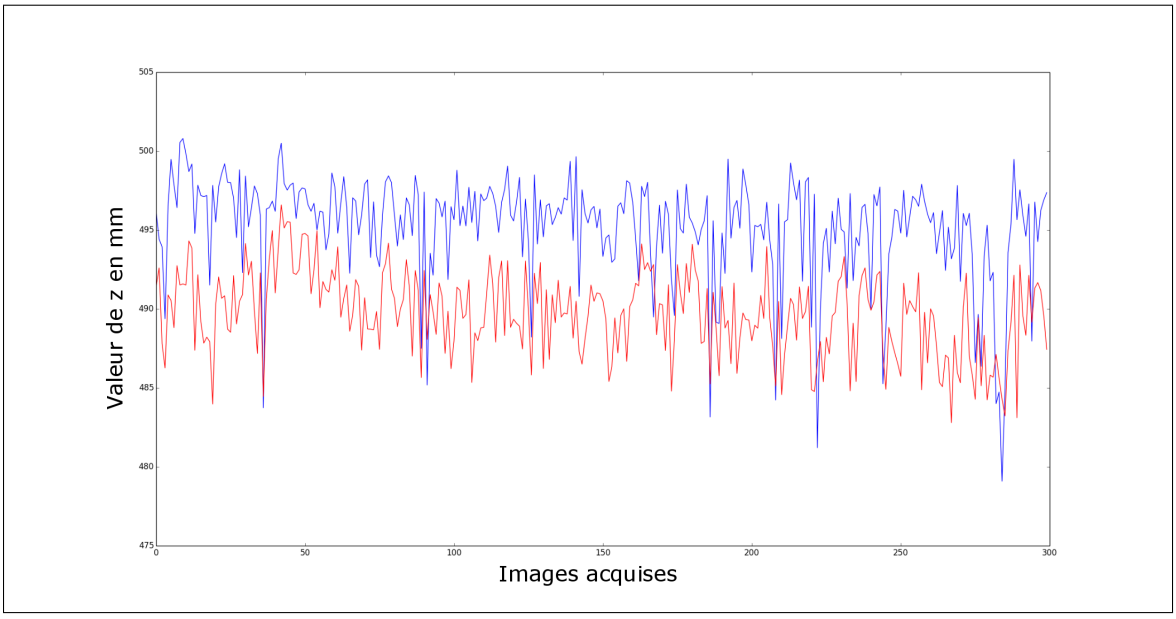
Un changement important de la valeur de z semble se dessiner avec le sujet 3, durant les 125 premières acquisitions, la moyenne est passablement plus basse que durant les suivantes, on peut donc deviner un mouvement comme un redressement du sujet par exemple. Le tracé rouge démontre bien la différence entre la valeur de z du sujet 3 par rapport aux deux autres. Ce qui indique une position de départ relativement plus près de la caméra comparés aux deux autres, qui peut être causé simplement par la posture du sujet ou par ses habitudes d'observations. On constate dans les trois sujets, un mouvement de la personne devant son écran. Ce qui fait que notre système de suivi du regard doit être invariant à ce mouvement. Ce que nous allons vérifier avec les expérimentations sur le suivi du regard.

On peut en conclure que la méthode avec marqueur est une méthode qui perd de la fiabilité plus la distance augmente entre la lentille et le point de la scène. Utiliser une

Sujet 1



Sujet 2



caméra avec une distance focale et une ouverture plus grande pourrait améliorer les résultats, mais il faut garder en tête l'objectif d'utiliser une caméra de type webcam comme celles intégrées dans les ordinateurs portables. Poursuivons maintenant avec

Sujet 3

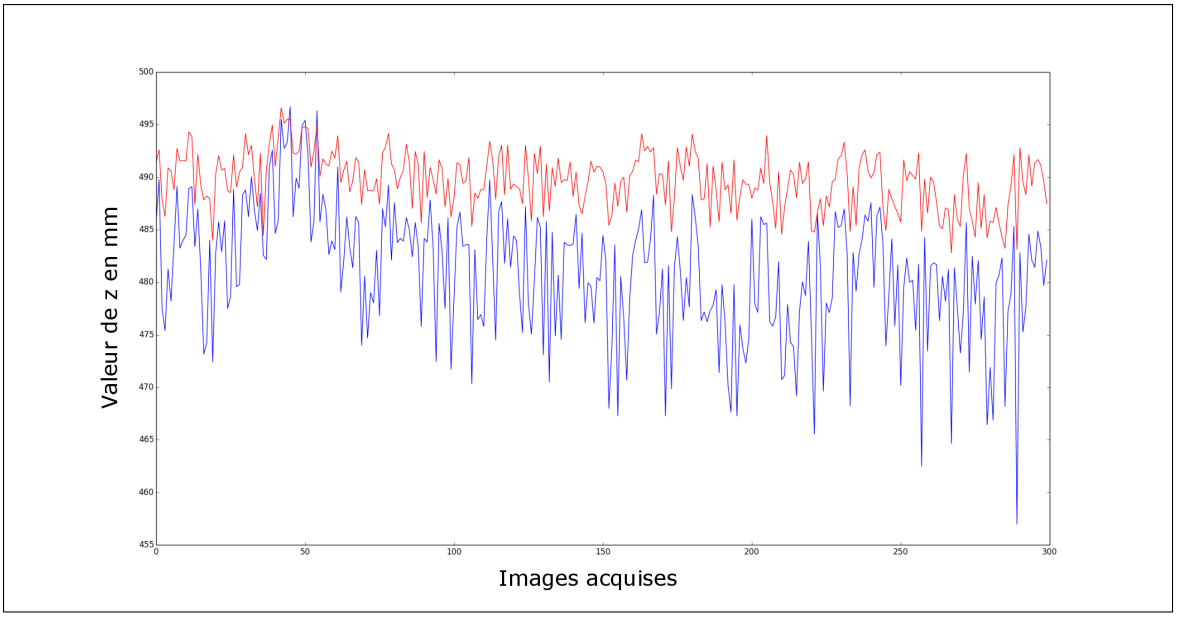


figure 4.1 – Représentation de z par rapport aux images acquises pour 3 sujets. Le tracé rouge représente la moyenne des 3 sujets

la même procédure, mais pour la méthode utilisant le paramètre du flou σ.

4.3 Profondeur z selon σ

Nous avons l'équation 2.16 du chapitre 2 donnant la profondeur z d'un point de la scène que nous avons retranscrite ici :

$$z = \begin{cases} \frac{fv}{v-f-kN\sigma} & \text{si z} > \text{u} \\ \frac{fv}{v-f+kN\sigma} & \text{si z} < \text{u} \end{cases}$$

À partir de cette équation nous pouvons tracer le graphique 4.2 correspondant à la profondeur z en mm en fonction de σ. Une représentation visuelle des résultats est également proposée sous forme de tableau 4.3. Rappelons que nous utilisons la distance focale $f = 4.033$ mm. De plus, puisque $u = 500$ mm nous utiliserons seulement la deuxième partie de l'équation ($z < u$) puisque les expérimentations se feront

toujours à une distance limite de 500 mm. Lié à l'ouverture N, pour que le flou reste proportionnel au disque d'un CdC, la constante $k = 0.00135$.

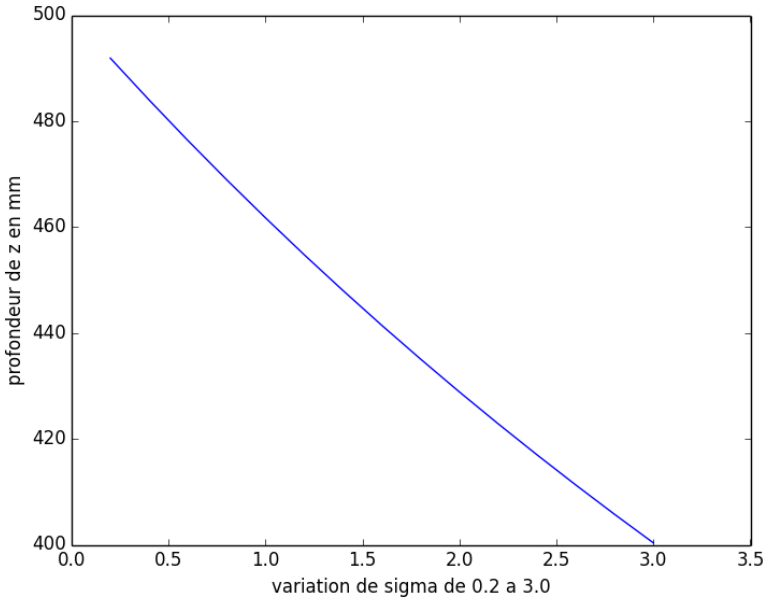


figure 4.2 – Courbe de z selon σ

σ	Z en mm
0.2	491.90
0.4	484
0.6	476.34
0.8	468.93
1.0	461.74
1.2	454.76
1.4	448
1.6	441.43
1.8	435.06
2.0	428.86
2.2	422.84
2.4	416.99
2.6	411.29
2.8	405.75
3.0	400.36

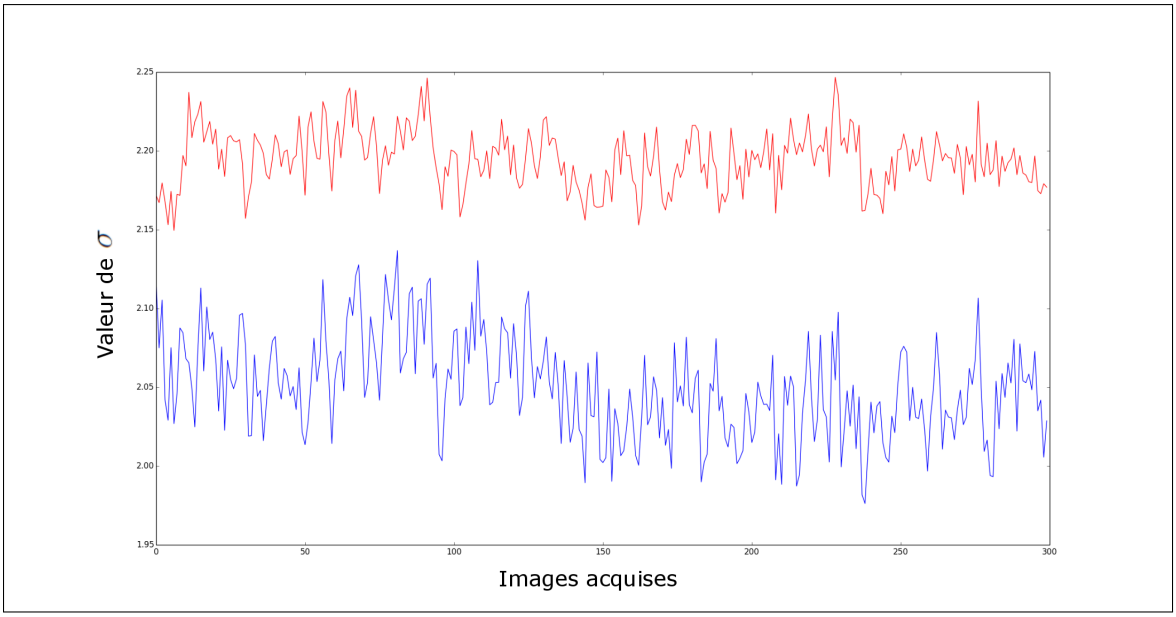
figure 4.3 – Profondeur Z selon σ

Le graphique et le tableau permettent de constater que plus σ est faible, z se rapproche de la distance au focus u. Ainsi, puisque l'on fait le calcul de σ pour chacune des images d'une séquence pour l'estimation du regard, il est important de vérifier la stabilité de σ dans nos images d'acquisitions comme nous avons fait pour le marqueur.

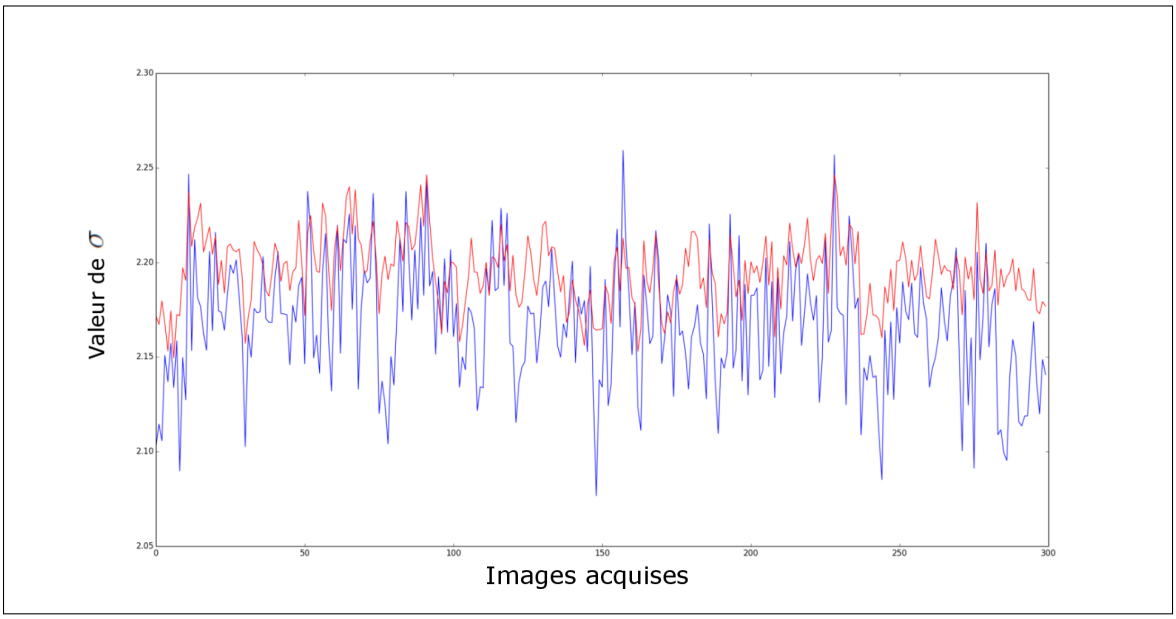
Encore une fois, nous avons fait trois nouveaux ensembles d'images provenant de 3 sujets différents. Cette fois-ci, nous avons suivi la valeur de σ pour chaque image acquise. Les trois ensembles possèdent 300 acquisitions. La figure 4.4 illustre la variation de σ et non la distance en mm. Le tracé rouge est la moyenne de sigma des trois ensembles.

Comme avec le marqueur, on remarque les mêmes phénomènes. La profondeur moyenne des sujets 2 et 3 est relativement constante. Celle de l'ensemble 1 varie

Sujet 1



Sujet 2



légèrement à partir de l'acquisition 150, ce qui laisse supposer un ajustement de la posture qui a provoqué un léger mouvement. La valeur de σ varie très peu, il est donc important de vérifier si cette estimation de σ est précise ou si elle fait défaut.

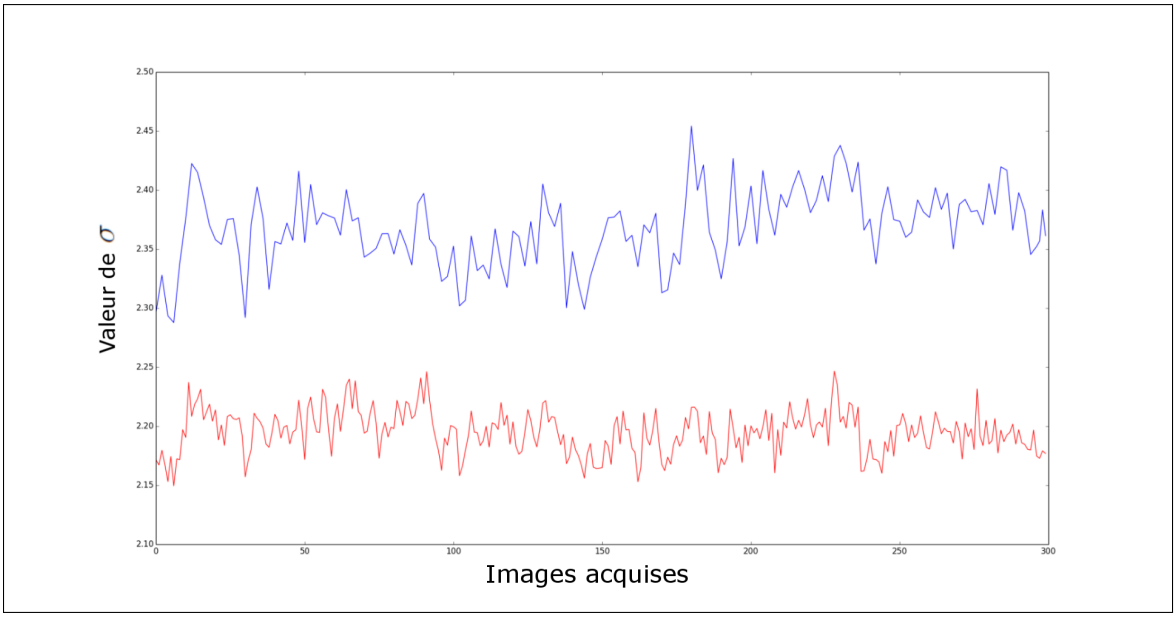


figure 4.4 – Représentation de σ par rapport aux images acquises pour trois sujets. Le tracé rouge représente la moyenne des 3 sujets

Sujet	Médianne	Moyenne σ	Minimum	Maximum	Variation σ	Variation de z en mm
1	2.0490	2.05	1.9763	2.1368	0.1605	4.8615
2	2.1684	2.1668	2.0767	2.2592	0.1825	5.4422
3	2.3696	2.3679	2.2879	2.4542	0.1663	4.8224

tableau 4.3 – Évaluation de σ à travers les images acquises

Par le tableau 4.3 on constate que la variabilité de z est moins grande que celle du marqueur. Cependant, la moyenne de σ pour chacun des sujets est trop élevée si nous voulons avoir une moyenne se rapprochant de 500 mm. L'ouverture de 2 mm de la caméra peut rendre la détection du flou moins précise part la petite quantité de rayons qui traverse la lentille. De plus, comme nous le verrons dans les expérimentations, le flou est calculé dans une région d'intérêt se situant sur le bout du nez. Il peut donc y avoir une variation supplémentaire d'une dizaine de millimètres à celle énoncée dans la section 4.2.

Nous pouvons conclure que la méthode avec marqueur fait une approximation de la profondeur d'un point de la scène avec une erreur d'estimation qui est plus faible que la méthode du flou. Le marqueur qui utilise un calcul direct de la profondeur avec un changement de repère est une méthode plus précise que la deuxième, mais elle reste une méthode encombrante et moins agréable pour l'utilisateur. Poursuivons avec l'analyse des résultats pour l'estimation du regard en parcourant différentes expérimentations.

4.4 Estimation du regard : Expérimentation 1

Pour la première expérimentation, nous avons limité une zone sur l'écran que l'usagé regardera. Cette zone est subdivisée en neuf rectangles. Les dimensions de l'écran sont respectivement pour la largeur et la hauteur de 52.5 x 57 cm. Chacun des rectangles a une taille de 15.5 x 19 cm. La figure 4.5 donne un aperçu de l'environnement de suivi du regard. Lors de l'acquisition des images qui serviront d'apprentissage, le sujet a comme consigne de regarder le centre de chacune des régions.

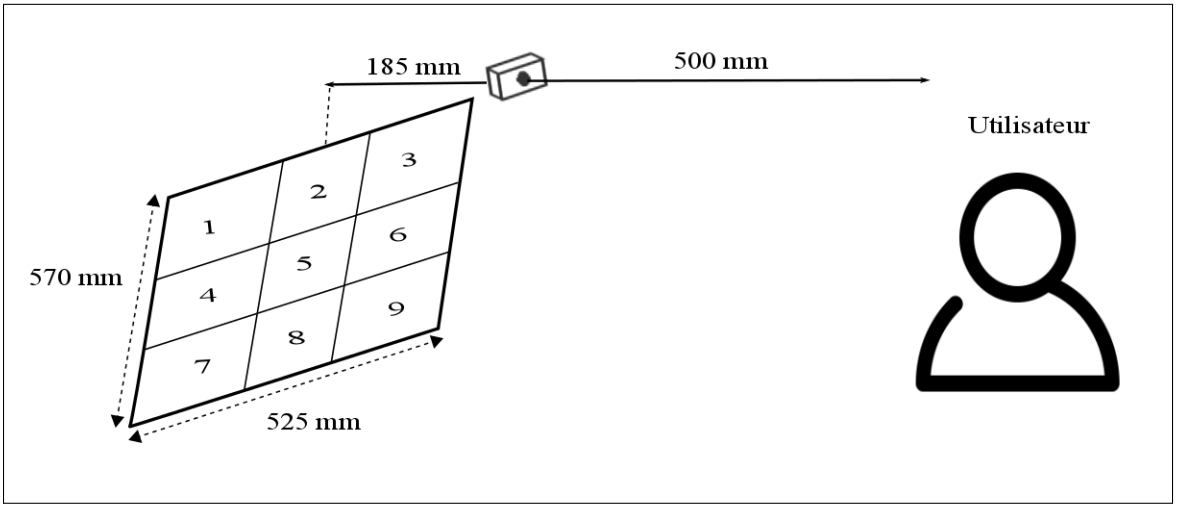


figure 4.5 – Environnement de suivi du regard lors l'expérimentation 1

La caméra d'acquisition est devant un écran à une distance d'environ 185 mm. L'utilisateur est placé à une distance maximale de 500 mm de la caméra pour une

distance totale entre l'utilisateur et l'écran d'observation oscillant autour de 685 mm. La distance provient du fait qu'il est conseillé par le Gouvernement du Canada, par l'entremise du Centre canadien d'hygiène et de sécurité au travail (CCHST) [3], d'être à une distance entre 400 et 740 mm de l'écran pour contrer la fatigue oculaire, c'est donc une distance réaliste.

Nous avons recueilli deux ensembles de 300 images pour une personne test, soit un ensemble pour la méthode du marqueur et un autre pour la méthode du flou. Après avoir enlevé les faux positifs, qui peuvent être causés par un clignement d'oeil ou le détournement du regard, nous avons pour chaque méthode, de façon aléatoire, formé deux ensembles d'entraînement formés à l'aide d'un modèle bayésien variationnel. Les ensembles d'entraînement servent à la formation du modèle statistique de la régression logistique multinomiale. Le premier ensemble d'entraînement est construit à partir de 100 images, puis nous avons ajouté 150 images pour former le deuxième ensemble d'entraînement de 250 images et le reste des images, servira pour l'ensemble de données tests. Nous avons formé deux ensembles de grosseurs différentes pour vérifier si l'augmentation de données dans un ensemble d'entraînement peut avoir un impact positif dans le pourcentage de classification. Les données tests sont les mêmes dans les deux cas. Le modèle statistique utilisé pour la phase d'entraînement sera le VBMLR [46] de même que pour la classification comme expliquée précédemmentdans la section 3.3. Remarquons que parfois nous avons moins que 50 données pour une région. Ceci est causé par les faux positifs que nous avons enlevés. On peut observer les résultats du pourcentage de classification par rapport à la région observée (RO) par l'utilisateur pour les deux méthodes dans les tableaux 4.4 à 4.7 suivants :

Méthode avec marqueur

RO \ Classification	1	2	3	4	5	6	7	8	9	%	#
1	**50**									100	50
2		**47**		3						94	50
3			**47**							100	47
4				**50**						100	50
5					**50**					100	50
6						**50**				100	50
7				1			**49**			98	50
8					6			**44**		88	50
9									**50**	100	50
Total										**97.7777**	**447**

tableau 4.4 – Apprentissage = 100 données, tests $\leq$ 50 données

RO \ Classification	1	2	3	4	5	6	7	8	9	%	#
1	**50**									100	50
2		**50**								100	50
3			**47**							100	47
4		1		**49**						98	50
5					**50**					100	50
6						**50**				100	50
7				1			**49**			98	50
8					4			**46**		92	50
9									**50**	100	50
Total										**98.6666**	**447**

tableau 4.5 – Apprentissage = 250 données, tests $\leq$ 50 données

À partir d'une collecte de données provenant d'une même personne, le VBMLR donne des scores élevés. Une plus grande quantité de données pour la phase d'entraînement a provoqué une amélioration des résultats pour les mêmes données tests. On souligne également que les deux méthodes ont des résultats similaires pour le score total. On peut cependant penser que la méthode du marqueur a un avantage minime car elle semble faire une meilleure distinction entre les régions contrairement à la méthode du flou semble avoir une confusion entre la région 8 et 9. La précision du marqueur dans le calcul de z mais également dans le calcul du positionnement spatial (R et T) peut peut-être expliquer cela. Cette expérimentation montre qu'une discrimination est possible pour le suivi du regard à l'aide du VBMLR et les vecteurs caractéristiques que nous avons définis. Refaisons maintenant la même expérimentation, mais dans un environnement différent pour confirmer notre analyse.

Méthode utilisant le flou

Classification / RO	1	2	3	4	5	6	7	8	9	%	#
1	**44**			4						91.6667	48
2		**39**								100	39
3			**49**							100	49
4				**48**						100	48
5					**50**					100	50
6					8	**42**				84	50
7							**50**			100	50
8								**49**	1	98	50
9								7	**43**	86	50
Total										**95.5185**	**434**

tableau 4.6 – Apprentissage = 100 données, tests $\leq$ 50 données

Classification / RO	1	2	3	4	5	6	7	8	9	%	#
1	**48**									100	48
2		**39**								100	39
3			**49**							100	49
4				**48**						100	48
5					**50**					100	50
6						**50**				100	50
7							**50**			100	50
8								**49**	1	98	50
9								6	**44**	88	50
Total										**98.4444**	**434**

tableau 4.7 – Apprentissage = 250 données, tests $\leq$ 50 données

4.5 Estimation du regard : Expérimentation 2

Dans cette expérimentation, nous avons changé la taille de l'écran et l'environnement de suivi du regard. L'écran a une diagonale de 17 pouces, soit un écran d'ordinateur portable standard. Nous avons donc un écran de 37 cm de largeur et de 23.5 cm de hauteur avec des rectangles d'observations d'une taille de 12.3 cm par 7.83 cm.

Cette fois l'écran et la caméra sont à la même distance et comme pour un ordinateur portable, la caméra est située juste au-dessus de l'écran 4.6. L'utilisateur est encore une fois à environ 500 mm de la caméra. Un sujet reçoit les mêmes indications que l'expérimentation 4.4. Cette fois, pour 2 sujets différents, nous avons recueilli deux ensembles d'images pour chaque méthode. Nous avons donc deux fois

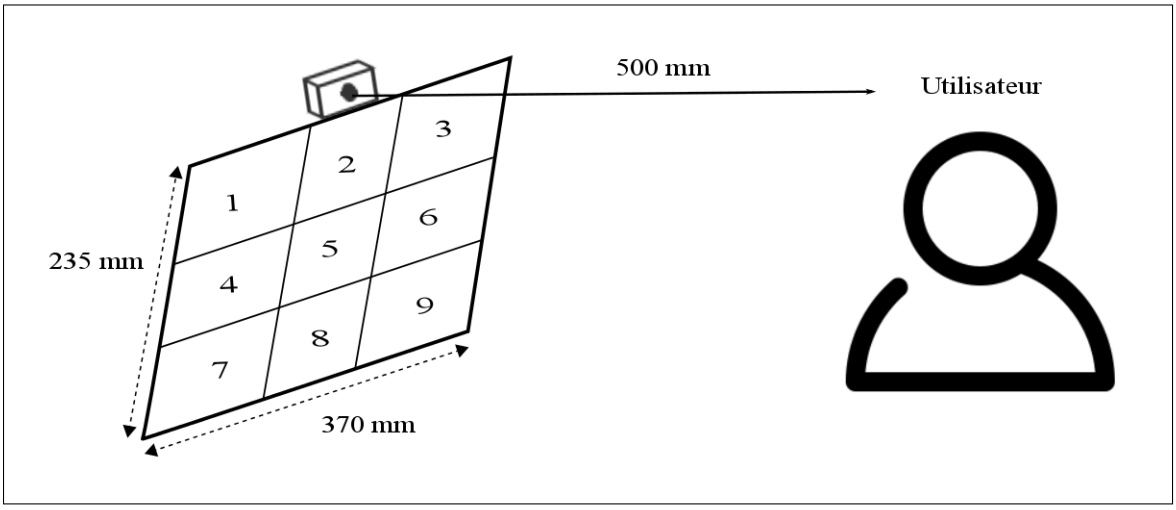


figure 4.6 – Environnement de suivi du regard lors des expérimentations 2 et 3

300 images pour chaque méthode et chaque sujet. Pour chaque ensemble de données, nous avons créé un ensemble d'entraînement de 250 images et 50 ont été utilisées pour le test de classification. Les ensembles ont été formés de façon aléatoire et le modèle statistique utilisé est le VBMLR. Les résultats sont représentés dans les tableaux 4.8, 4.9, 4.10 et 4.11.

Méthode avec marqueur

RO \ Classification	1	2	3	4	5	6	7	8	9	%	#
1	**36**	6		8						72	50
2	2	**47**								94	50
3			**49**			1				98	50
4	16	2		**31**	1					62	50
5		4			**41**		5			100	50
6						**50**				100	50
7					4		**43**	3		86	50
8								**50**		100	50
9									**50**	100	50
Total										**88.2222**	**450**

tableau 4.8 – Sujet 1 : apprentissage = 250 données, test $\leq$ 50 données

Classification / RO	1	2	3	4	5	6	7	8	9	%	#
1	**48**									100	48
2		**48**	1							97.9592	49
3			**50**							100	50
4		2		**37**	10					75.5102	50
5					**50**					100	49
6						**50**				100	50
7							**49**	1		98	50
8								**50**		100	50
9									**50**	100	50
Total										**98.6666**	**446**

tableau 4.9 – Sujet 2 : apprentissage = 250 données, test $\leq$ 50 données

Méthode utilisant le flou

Classification / RO	1	2	3	4	5	6	7	8	9	%	#
1	**47**									100	47
2		**50**								100	50
3			**48**							100	48
4	46			**4**						8	50
5					**48**				2	96	50
6						**42**			8	84	50
7							**17**	33		34	50
8								**40**	1	97.561	41
9								12	**38**	76	50
Total										**77.2844**	**436**

tableau 4.10 – Sujet 1 : apprentissage = 250 données, tests $\leq$ 50 données

RO \ Classification	1	2	3	4	5	6	7	8	9	%	#
1	**50**									100	50
2		**48**	1							97.9592	49
3			**49**							100	49
4				**36**	3		8			76.5957	47
5					**31**		18			63.2653	49
6						**45**	3		1	91.8367	49
7					11	3	**36**			72	50
8								**37**	13	74	50
9								4	**46**	92	50
Total										**85.2952**	**443**

tableau 4.11 – Sujet 2 : apprentissage = 250 données, tests $\leq$ 50 données

Pour les deux méthodes, les résultats de l'expérimentation 2 subissent une baisse considérable par rapport à la première expérimentation excepté pour l'ensemble 2 de la méthode du marqueur. La méthode avec marqueur reste la plus précise et comme mentionné préalablement, elle est très sensible au positionnement spatial contrairement à la méthode utilisant le flou qui trace un vecteur de direction par rapport à la détection du nez et du centre de la tête. Ce qui peut expliquer la différence des performances entre les deux méthodes. La diminution de la taille des régions apporte de la confusion entre certaines régions collées. Par exemple, dans l'ensemble 1 de la méthode du flou (tableau 4.10), une grande confusion entre les régions 4 et 1 existe, de même qu'avec les régions 7 et 8, de sorte que le score de la classification finale est grandement diminué. Cette confusion est aussi expliquée par le fait que pour certain sujet, ce sont les iris qui jouent le plus grand rôle lorsque s'agit d'avoir une observation plus précise. Le vecteur de disparité n'arrive donc pas à discriminer clairement certaines régions. Bien que pour la majorité des régions la classification est encore au-delà de 95 %, une discrimination adéquate pour l'ensemble des régions est difficile à atteindre pour une surface telle qu'un écran d'ordinateur.

Nous voulons cependant que notre méthode soit invariante à la personne qui l'utilise. La méthode avec marqueur a été implémentée principalement pour faire la transition entre une méthode à plusieurs caméras et une méthode à une seule caméra. Par la suite, l'estimation du flou nous a permis de calculer la profondeur et de laisser tomber tout artifice lié à un élément externe, comme un marqueur. Nous allons donc laisser tomber pour la dernière expérimentation la méthode du marqueur. De plus,

lors de l'expérimentation 1, les résultats étaient satisfaisants pour les deux méthodes. Il serait également difficile de placer le marqueur à la même position sur chacun des sujets vu la physionomie des visages. L'orientation étant très sensible, un marqueur sur un front plus prononcé vient changer les caractéristiques de directions ce qui rend l'invariance entre les sujets difficile pour cette méthode.

4.6 Estimation du regard : Expérimentation 3

Dans cette expérimentation nous avons voulu vérifier l'invariance du système pour différentes personnes. À cause de la confusion trop grande entre certaines régions lors de l'expérimentation 2, nous avons repris l'environnement de l'expérimentation 1. Faisant partie de l'expérimentation 1 et étant un des deux sujets de l'expérimentation 2, nous avons ajouté 3 nouvelles personnes dans la collection d'images avec moi comme sujet. L'autre sujet de l'expérimentation 2 n'étant pas disponible ne se retrouve donc pas dans l'expérimentation 3. Lors de l'acquisition des données nous avons simplement placé le sujet dans l'environnement en lui soumettant comme seule instruction de fixer la région énoncée de la façon la plus confortable et naturelle possible. La collection des images se retrouvera finalement à être notre ensemble d'entraînement par l'extraction que nous pouvons faire des vecteurs de caractéristiques de chaque personne. L'hypothèse est que cette procédure permet d'inclure un plus large éventail des habitudes d'observation puisque chaque personne peut avoir une manière précise d'observer une scène comme nous en avons fait la mention dans le chapitre 3 courant.

Pour le premier test, nous avons fait l'acquisition de 300 images par personne. Pour chaque sujet, nous avons de manière aléatoire, extrait 150 images pour chaque région d'observation. Nous avons donc formé un ensemble entraînement de 4 x 150 images pour chacune des régions. Ce qui nous donne un ensemble composé de 4 personnes différentes et un total de 600 images pour chaque région. Par la suite, dans les images restantes pour chaque sujet, nous avons pris un échantillon de 75 images par région comme ensemble de validation. Encore une fois de manière aléatoire. Le sujet est une des quatre personnes de la collection et les pourcentages de classification sont présentés dans le tableau 4.12.

Test 1 : Méthode du flou avec 4 sujets

Classification / RO	1	2	3	4	5	6	7	8	9	%	#
1	**27**	48								36	75
2		**68**	7							90.6667	75
3			**75**							100	75
4	3			**66**		4	2			88	75
5					**71**	4				94.6667	75
6						**75**				100	75
7				9			**66**			88	75
8								**71**	4	94.6667	75
9									**75**	100	75
Total										**88**	**675**

tableau 4.12 – apprentissage = 600 données, tests = 75 données

On peut observer une diminution du pourcentage de classification totale. Ce qui est normal puisque l'ajout de différentes personnes peut provoquer un élargissement ou un rétrécissement de la densité de probabilité de certaines régions. Une région peut contenir des vecteurs caractéristiques ayant une différence plus marquée que d'autres. Un sujet peut par exemple orienter la tête de façon plus prononcée comparativement à un autre pour les régions se situant plus basse à cause de lunettes. Poursuivons maintenant avec le même test, mais pour un sujet n'étant pas dans la collection d'images. Nous avons donc un ensemble d'apprentissages composé de 3 x 150 données. Les résultats se retrouvent encore une fois dans le tableau 4.13.

On observe dans le dernier résultat, une baisse du taux de bonne classification. Malgré le fait que certains scores soient élevés pour des régions, on remarque surtout une confusion entre les régions 1, 5, 6 et 7. On pense q'un plus grand échantillon de personnes pourrait rectifier cela. Pour ces régions ayant un score de zéro, une supposition se fait au niveau d'une trop grande différences entre les vecteurs de caractéristiques. C'est-à-dire qu'une région qui a des vecteurs de caractéristiques trop différents entre les 3 sujets, aurait une densité de probabilité trop faible par rapport aux autres régions dans l'ensemble d'entraînement. Concrètement, on suppose donc que les sujets peuvent avoir des manières de regarder qui diffèrent trop des uns des autres. Par exemple, un sujet peut pencher la tête de façon plus prononcée avant

Test 2 : Méthode du flou avec un sujet n'étant pas dans la collection d'images

Classification / RO	1	2	3	4	5	6	7	8	9	%	#
1	**1**	17	57							1.3333	75
2	3	**72**								96	75
3			**75**							100	75
4				**75**						100	75
5	20	44	1	1	**0**	9				0	75
6			75			**0**				0	75
7				74			**1**			1.3333	75
8								**75**		100	75
9								1	**74**	98.6667	75
Total										**55.2593**	**675**

tableau 4.13 – apprentissage = 450 données, tests = 75 données

d'ajuster son regard vers la région voulu avec ses iris, tandis qu'un autre sujet pourrait seulement bouger les iris et que très peu la tête. La solution, qui serait à confirmer, serait d'augmenter le nombre de sujets dans la collection d'images pour inclure les différentes habitudes d'observation et consolider ainsi la discrimination des régions. Poursuivons maintenant avec l'évaluation des performances avant de conclure avec une synthèse des résultats.

4.7 Estimation du regard : Temps de calcul des performances

Pour l'évaluation des performances, nous avons exécuté l'application et fait une moyenne du temps de calcul pour l'estimation du regard pour 100 images. La caméra produit 20 images par seconde. L'application est entièrement en langage C++ sous une interface Qt. L'implémentation de l'interface nécessite la conversion d'image en un format différent que celui d'acquisition soit, le format QImage, pour pouvoir en faire l'affichage. Cette conversion demande un temps de calcul supplémentaire. L'application complète : capturer l'image d'acquisition, la création d'un vecteur de caractéristiques, jusqu'à l'identification de la région du regard, demande en moyenne 0.51468 seconde par image. Ce qui signifie que le suivi du regard se fait chaque demi-seconde. Ce qui n'est pas suffisant pour être en temps réel. Les différentes étapes ainsi

que leurs temps moyens sont affichés dans le tableau 4.14.

Étape	Nombre d'images par seconde
Détection centre de l'iris	3.94
Calcul du flou	9.23
vecteurs R et T	925.93
vecteur de disparité	1086.96
Classification	52.63
Application complète	1.94

tableau 4.14 – Évaluation des performances des différentes étapes de l'application

La détection du centre de l'iris est l'étape la plus lourde en temps de calcul et celle-ci demeure problématique en plus de l'étape du calcul du flou. L'optimisation de ses étapes de même que de l'interface est nécessaire pour atteindre le temps réel. Cependant, une parallélisation est faite pour pouvoir utiliser l'application en temps réel. L'acquisition des images et la construction d'un vecteur caractéristique de même que la classification du regard sont faits de manière à ne pas s'obstruer. C'est-à-dire que la caméra de l'application fonctionne en temps réel en parallèle avec les étapes permettant de faire la classification du regard. Le système fait une mise à jour du regard 1.94 fois par seconde pour affirmer dans quelle région le regard est rendu. Si nous voulons atteindre le temps réel, la rapidité du calcul de détection du centre de l'iris est donc la première étape en perspective d'amélioration du système. Il en est de même pour l'étape du calcul du flou.

4.8 Conclusion

Pour conclure ce chapitre sur les expérimentations et les résultats concernant le suivi du regard, nous avons pu prendre conscience de la complexité que représente un système permettant de faire le suivi du regard. Au travers la littérature, on remarque, que plusieurs choix s'imposent : un système à une ou plusieurs caméras, la méthode utilisée pour la détection des différentes parties, un modèle statistique, etc. Nous avons présenté dans ce chapitre deux méthodes d'estimation du suivi du regard. La méthode

avec marqueur conduit à un suivi du regard plus précis. L'inconvénient réside dans la pose du marqueur sur l'individu qui rend l'utilisation moins agréable. Elle empêche également de construire un système invariant à la personne. La méthode utilisant le flou est mieux adaptée dans le sens qu'elle demeure une méthode novatrice avec aucun calibrage. L'inconvénient demeure que pour atteindre un standard de performance élevé pour toute personne voulant utiliser le système, la construction d'une collection d'images contenant beaucoup plus de sujets est nécessaire. Nous croyons que de cette manière, la méthode du flou deviendrait performante peu importe si l'utilisateur se retrouve dans la collection ou non puisqu'un système semblable au point de vue statistique a déjà été tenté dans [20]. Il reste également à innover dans la façon de l'implémenter pour produire un système temps-réel.

Conclusion

Au cours de cette maîtrise, à travers le cheminement d'un rayon lumineux dans le processus de formation d'image, nous avons introduit les différents concepts et facteurs qui peuvent influencer la formation de l'image d'un système optique. Que ce soit par la luminosité de la scène, la nature de l'optique ou les caractéristiques du capteur, nous avons vu qu'une perte d'information est incontournable. De même qu'un amoncellement d'une quantité de flou dans les images lors de l'arrivée des rayons sur le capteur. Le flou représente un phénomène que nous avons étudié par ses concepts et les terminologies l'entourant. Il est directement lié au processus de formation d'image et à la profondeur d'un objet par rapport à la lentille d'une caméra. Mathématiquement, les cercles de confusion peuvent être simulés par des fonctions de points d'étalement. Cette mesure de flou a permis par la suite de créer un système d'estimation du regard. Une plateforme a été implémentée avec l'intégration d'un système du suivi du regard existant [20]. Puis, deux autres systèmes, que nous avons implémentés, ont été ajoutés. Le premier est une méthode avec marqueur et le deuxième, utilisant le flou. La grande spécificité de ces méthodes est le fait de n'utiliser qu'une seule caméra. Mais l'apport principal de cette maîtrise est d'avoir intégré le flou d'une image acquise pour faire le suivi du regard d'un utilisateur. Dans la littérature, aucune méthode n'en fait mention. Au final, c'est donc un système fonctionnant avec une seule caméra de type webcam, mais également sans aucune calibration externe à la caméra. Une méthode originale avec des résultats positifs pour un utilisateur se trouvant dans la collection d'images. La prochaine étape serait donc d'enrichir la collection en trouvant de nouveaux sujets et en élargissant la taille de l'échantillonnage de chacun des sujets. L'élaboration d'une nouvelle méthode de détection du centre de l'iris est également dans les plans pour rendre le système temps réel. Par la suite, l'objectif serait de pouvoir subdiviser en

un plus grand nombre de régions l'écran pour améliorer la précision du regard.

Bibliographie

[1] David Beymer et Myron Flickner.
« Eye gaze tracking using an active stereo head ».
Dans *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–451–8 vol.2, June 2003.

[2] Xavier L. C. Brolly et Jeffrey B. Mulligan.
« Implicit Calibration of a Remote Gaze Tracker ».
Dans *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, pages 134–134, June 2004.

[3] CCHST.
« CCHST : Centre canadien d'hygiène et de sécurité au travail ».
http://www.cchst.ca/oshanswers/ergonomics/office/monitor_positioning.html.

[4] Robert L. Cook et Kenneth E. Torrance.
« A Reflectance Model for Computer Graphics ».
ACM Trans. Graph., 1(1):7–24, janvier 1982.

[5] Julien Couillaud.
« Formation d'image : Estimation du champ lumineux et matrice de filtres couleurs ».
Mémoire de maîtrise, Université de Sherbrooke, Département d'informatique, Sherbrooke, Québec, Canada, 2012.

[6] Julien Couillaud, Alain Horé et Djemel Ziou.
« Nature-inspired color-filter array for enhancing the quality of images ».
J. Opt. Soc. Am. A, 29(8):1580–1587, Aug 2012.

[7] François Deschênes.
« *Estimation des jonctions, du mouvement apparent et du relief en vue d'une reconstruction 3D de la scène* ».
Thèse de doctorat, Département d'informatique, Université de Sherbrooke, Sherbrooke, Québec, Canada, 2006.

[8] François Deschenes, Djemel Ziou et Philippe Fuchs.
« Improved Estimation of Defocus Blur and Spatial Shifts in Spatial Domain : A Homotopy-Based Approah ».
Pattern Recognition, 36(9):2105–2125, 2003.

[9] François Deschenes, Djemel Ziou et Philippe Fuchs.
« An Unified Approach for a Simultaneous and Cooperative Estimation of Defocus Blur and Spatial Shifts ».
Image and Vision Computing, 22(11):35–57, 2004.

[10] Andrew T. Duchowski.
Eye Tracking Methodology : Theory and Practice.
Springer-Verlag New York, Inc., 2007.

[11] David A. Forsyth et Jean Ponce.
Computer Vision : A Modern Approach.
Prentice Hall Professional Technical Reference, 2002.

[12] Natalie L. Garrett.
« Introduction to Physics in Modern Medicine, 2nd edn., by Suzanne Amador Kane ».
Contemporary Physics, 52(2):164–165, 2011.

[13] Craig Hennessey, Borna Noureddin et Peter Lawrence.
« A Single Camera Eye-gaze Tracking System with Free Head Motion ».
pages 87–94, 2006.

[14] Alain Hore et Djemel Ziou.
« An Edge-Sensing Generic Demosaicing Algorithm With Application to Image Resampling ».
IEEE Transactions on Image Processing, 20(11):3136–3150, Nov 2011.

[15] Akihiro Horii.
« Depth from Defocusing ».
Rapport Technique, 1992.

[16] Berthold Klaus Paul Horn, éditeur.
Robot Vision.
MIT Press, 1986.

[17] Wen-Bing Horng, Chih-Yuan Chen, Yi Chang et Chun-Hai Fan.
« Driver fatigue detection based on eye tracking and dynamk, template matching ».
Dans *IEEE International Conference on Networking, Sensing and Control, 2004*, volume 1, pages 7–12, March 2004.

[18] Vicon Industries.
Megapixel camera in security application : Are they ready for prime time ?
Rapport Technique, Vicon Industries, January 2009.

[19] irmquebec.
« IRM Québec, Imagerie par résonance magnétique ».
https://irmquebec.com/lirm/appareils/.

[20] Reza Jafari et Djemel Ziou.
« Gaze estimation using Kinect/PTZ camera ».
Dans *ROSE 2012 : Magdeburg, Germany*, pages 13–18, 2012.

[21] Qiang Ji et Xiaojie Yang.
« Real-Time Eye, Gaze, and Face Pose Tracking for Monitoring Driver Vigilance ».
Real-Time Imaging, 8(5):357 – 377, 2002.

[22] Anat Levin, Rob Fergus, Frédo Durand et William T. Freeman.
« Image and Depth from a Conventional Camera with a Coded Aperture ».
ACM Trans. Graph., 26(3), juillet 2007.

[23] Margaret Livingstone.
Vision and art : the biology of seeing.
Harry N. Abrams, 2002.

[24] André Meyer, Martin Böhme, Thomas Martinetz et Erhardt Barth.
« A Single-Camera Remote Eye Tracker ».
Dans *Perception and Interactive Technologies : International Tutorial and Research Workshop, PIT 2006 Kloster Irsee, Germany, June 19-21, 2006. Proceedings*, pages 208–211.
Springer Berlin Heidelberg, 2006.

[25] Takehiko Ohno, Naoki Mukawa et Atsushi Yoshikawa.
« FreeGaze : A Gaze Tracking System for Everyday Gaze Interaction ».
Dans *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, ETRA '02, pages 125–132. ACM, 2002.

[26] OpenCV.
« OpenCV documentation, Camera calibration With OpenCV ».

[27] Rik Pieters, Edward Rosbergen et Michel Wedel.
« Visual Attention to Repeated Print Advertising : A Test of Scanpath Theory ».
Journal of Marketing Research, 36(4):424–438, 1999.

[28] Rajeev Ramanath, Wesley E. Snyder, Youngjun Yoo et Mark S. Drew.
« Color image processing pipeline ».
IEEE Signal Processing Magazine, 22(1):34–43, Jan 2005.

[29] Austin Roorda.
« Human visual system—image formation ».
Encyclopedia of imaging science and technology, 2002.

[30] Jens Rydell et Johan Eklöf.
« Vision complements echolocation in an aerial-hawking bat ».
Naturwissenschaften, 90(10):481–483, 2003.

[31] Erdmann Fred Schubert.
Light-Emitting Diodes.
Cambridge University Press, 2 édition, 006 2006.

[32] Sheng-Wen Shih et Jin Liu.
« A novel approach to 3-D gaze tracking using stereo cameras ».
IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 34(1):234–245, Feb 2004.

[33] Andor Technology.
« Digital Camera Fundamentals ».
Avril 2012.
http://www.lot-oriel.com/files/downloads/andor/en/cc_digitalcameras_deen01.pdf.

[34] Geoffrey Underwood.
« Eye Guidance in Reading and Scene Perception ».
pages v – vi. Elsevier Science Ltd, 1998.

[35] Geoffrey Underwood.
« Cognitive Processes in Eye Guidance : Algorithms for Attention in Image Processing ».
Cognitive Computation, 1(1):64–76, 2009.

[36] vetofish.
« vetofish spectre lumineux ».
https://www.vetofish.com/definition/spectre-lumineux.

[37] Vivotek.
The latest advances in megapixel surveillance.
Rapport Technique, July 2012.

[38] Michel Wedel et Rik Pieters.
« Eye Fixations on Advertisements and Memory for Brands : A Model and Findings ».
Marketing Science, 19(4):297–312, novembre 2000.

[39] R. G. White et R. A. Schowengerdt.
« Effect of point-spread functions on precision edge measurement ».
Journal of the Optical Society of America A, 11:2593–2603, oct 1994.

[40] Kurt Bernard Wolf et Guillermo Krotzsch.
« Geometry and dynamics in refracting systems ».
European Journal of Physics, 16(1):14, 1995.

[41] Günther Wyszecki et Walter Stanley Stiles.
« Color Science, Concepts and Methods. Quantitative Data and Formulas ».
Physics Bulletin, 18(10):353, 1967.

[42] Yalin Xiong et Steven Shafer.
« Depth from Focusing and Defocusing ».
Rapport Technique CMU-RI-TR-93-07, Robotics Institute, Pittsburgh, PA, March 1993.

[43] Zhiwei Zhu, Qiang Ji et Kristin P. Bennett.
« Nonlinear Eye Gaze Mapping Function Estimation via Support Vector Regression ».
Dans *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1132–1135, 2006.

[44] Shaojie Zhuo et Terence Sim.
« Defocus map estimation from a single image. ».
Pattern Recognition, 44(9):1852–1858, 2011.

[45] Djemel Ziou et François Deschênes.
« Depth From Defocus Estimation in Spatial Domain ».
Computer Vision and Image Understanding, 81:143–165, 1999.

[46] Djemel Ziou et Reza Jafari.
« Efficient steganalysis of images : learning is good for anticipation ».
Pattern Analysis and Applications, 17(2):279–289, 2014.

[47] Djemel Ziou, Shengrui Wang et Jean Vaillancourt.
« Depth from Defocus using the Hermite Transform ».
Dans *Proceedings of the 5th IEEE International Conference on Image Processing (ICIP'98)*, volume 2, page 958–962. IEEE Computer Society, IEEE Computer Society, 1998.