Estimation de la pose du visage du conducteur par caméra thermique

Samuel Bole

N° d'ordre NNT : 2017LYSES016

THÈSE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de

Université Jean-Monnet (Saint-Étienne)

Laboratoire Hubert Curien

École Doctorale N° 488

ED SIS Science Ingénérie Santé

Spécialité de doctorat : image, vision

Soutenue publiquement le 30/06/2017, par :

Samuel Bôle

Estimation de la pose du visage du conducteur par caméra thermique

---

Devant le jury composé de :

Tapus, Adriana	grade/qualité	établissement/entreprise	Président.e
Chausse, Frédéric	Maître de Conférences HDR	Université d'Auvergne	Rapporteur
Gravrand, Olivier	Directeur de recherche HDR	CEA Leti	Rapporteur
Nom, prénom	grade/qualité	établissement/entreprise	Examinateur.rice
Nom, prénom	grade/qualité	établissement/entreprise	Examinateur.rice
Lépine, Thierry	Maître de Conférences HDR	IOGS	Directeur de thèse
Fournier, Corinne	Maître de Conférences	Univeristé Jean Monnet	Co-directrice de thèse
Nom, prénom	grade/qualité	établissement/entreprise	Invité.e
Lavergne, Christophe	Référent optronique	Renault	Co-encadrant
Druart, Guillaume	Chargé de recherche	ONERA	Co-encadrant

Remerciements

Je remercie en premier lieu Frédéric Chausse et Olivier Gravrand pour accepté d'être mes rapporteurs.

Il me tient à cœur de remercier Corinne Fournier pour son investissement dans la co-direction et l'encadrement de cette thèse. Malgré la distance entre le site Renault de Guyancourt et le laboratoire Hubert Curien de Saint-Etienne, elle a su adapter son encadrement et rester très disponible. Je la remercie également d'avoir partagé, toujours avec pédagogie et bienveillance, une partie de ses connaissances pour me faire progresser. Enfin je la remercie pour les possibilités d'encadrement de stages de deuxième année à Télécom Saint-Etienne qu'elle m'a offerte. Les co-encadrements de stagiaires, que nous avons mené à deux, m'ont permis de progresser et de prendre du recul sur le sujet. J'en profite pour remercier ces trois stagiaires. Tout d'abord Jérôme Laporte, Jiyong Han qui ont travaillé en binôme durant l'été 2016. Ensuite Damien Joubert, qui a effectué son stage pendant l'été 2015. Merci à vous trois pour avoir supporté la chaleur caniculaire des salles vitrées du LaHC, et d'avoir implémenté et testé avec plus de rigueur que moi certaines des idées présentes dans ce manuscrit. Je vous souhaite à tous les trois une bonne continuation, et plus particulièrement une excellente thèse à Damien au sein de Renault.

Je remercie Guillaume Druart pour son enthousiasme communicatif, ainsi que pour les conseils qu'il m'a prodigués et les différentes relectures qu'il a pu mener. Son encadrement a été précieux pour découvrir de manière accélérée le domaine de l'infrarouge. J'adresse mes remerciements à Mathieu Chambon qui m'a permis de réaliser un certain nombre d'expérimentations avec des caméras thermiques de petit format. Ces expériences n'apparaissent pas dans ce manuscrit. Néanmoins, j'ai pu effectuer mes premiers étalonnages de caméras thermiques ainsi que des images à l'intérieur de l'habitacle grâce à son aide et au matériel de l'ONERA.

Je remercie également Thierry Lépine pour avoir dirigé ces travaux de thèse. Nos rencontres à Saint-Etienne furent toujours très agréables. Il a également mis à ma disposition du matériel en début de thèse, ce qui a orienté le choix du matériel commandé par la suite par Renault pour ce sujet.

Je tiens évidement à remercier Christophe Lavergne sans qui ce sujet de thèse n'aurait vu le jour. Malgré une charge importante de travail, ses convictions l'amènent à proposer des sujets qui vont au-delà des priorités essentielles et à court terme d'un industriel. J'ai pris beaucoup de plaisirs à travailler sur ce sujet passionnant et j'espère sincèrement qu'il continuera de vivre chez Renault.

J'aimerais remercier Erwan Aotore, Alain Durand et Emmanuel Bercier de la société *Ulis* pour les échanges que nous avons eus au cours de ces trois années autour des applications automobiles des imageurs microbolomètres. Leurs explications sur le fonctionnement de leurs détecteurs, toujours très claires, nous auront permis de monter en compétences et ainsi de poser quelques briques élémentaires pour l'analyse de faisabilité de la surveillance de l'habitacle par caméra thermique.

Je tiens à remercier les personnes de Renault qui gravitent autour du sujet *driver monitoring*, dont Stéphane Lallement et Philippe Saint-Loup (et désormais Joseph Morel) pour avoir partagé l'état de leurs connaissances sur le sujet. Je remercie également Philippe Labrevois pour son aide sur SolidWorks qui a consistée à compléter le maillage 3D du visage obtenue avec un scanner. Merci à lui pour son soutien sur RT-Maps ainsi que pour la réalisation du modèle simplifié du véhicule. Merci de m'avoir mis en relation avec Thierry Hodiesne pour occuper son atelier pendant les acquisitions des vidéos thermiques.

Je tiens à remercier Raymond Clerc pour son soutien technique sur les essais en enceinte climatique. Sa connaissance du fonctionnement de l'enceinte m'a été très utile. Raymond a également pu me conseiller sur la tenue aux conditions climatiques des pièces automobiles de l'habitacle.

Je tiens à remercier Olivier Bailly pour ses qualités managériales qui m'ont permis d'évoluer en toute sérénité à Renault. Son aide (ainsi que celle de Christophe), fut précieuse lors des phases d'achat du matériel. J'aimerais également remercier tous les membres permanents de l'équipe *ADAS amont manœuvre & environnement*. Votre bonne humeur m'a été précieuse pendant ces trois années et plus particulièrement pendant la rédaction de ce manuscrit. J'espère que nos routes professionnelles se croiseront à nouveaux. Je remercie toutes les personnes de Renault, permanentes ou non, qui ont accepté d'être filmées par une caméra thermique avec un bandeau sur le front. Bien que vous soyez tous très très beaux, je vous rassure, vous n'apparaissez pas dans le manuscrit.

J'adresse mes remerciements aux permanents, prestataires, stagiaires, apprentis et thésard du « labo opto » de Renault pour les bons moments et les coups de mains : Adrien, Jordi, Louis, Camille, Shuxian, Damien, Solal, Leana et Anastasia. Merci à Jordi, Louis, Camille et Solal d'avoir désembourbé ma voiture une journée d'hiver sur les bordures en herbe de « la fausse aux Loups ».

Enfin j'aimerais remercier ma famille et mes proches pour leur soutien. Je remercie en particulier ma mère pour sa relecture. Je remercie également Annaïg de m'avoir soutenu pendant ces trois années et plus particulièrement d'avoir contenu ses émotions encore six longs mois après sa soutenance de thèse.

Résumé

Les travaux de recherche présentés dans ce mémoire portent sur l'estimation de la pose du visage d'un conducteur par caméra thermique. Les applications potentielles sont : l'estimation de la direction du regard (pour estimer l'attention à la tâche de conduite), l'optimisation du déclenchement des airbags, l'adaptation des affichages 'tête haute'ou encore le contrôle de paramètres physiologiques (température, rythme respiratoire)…

Des systèmes actuellement commercialisés sont basés sur une caméra et un illuminant proche infrarouge (LEDs à ~ 900 nm). Le confort oculaire, sur le long terme, n'étant pas clairement établi, nous souhaitons développer une alternative à ces systèmes. Les caméras thermiques (bande 8-14 µm) en représentent une puisqu'elles sont particulièrement sensibles au rayonnement propre des objets à 30°C tels que les visages humains. De plus, les avancées technologiques autour de ces imageurs vont dans le sens d'une réduction des coûts pour adresser les marchés de grand volume tels que la domotique et l'automobile.

Dans le cadre de cette thèse, nous avons choisi d'utiliser une caméra thermique commerciale à microbolomètres. L'objet de ces travaux est de maîtriser les étapes de correction du Bruit Spatial Fixe (*BSF)* et le traitement de l'image afin d'estimer la pose du visage d'un conducteur en mettant l'accent sur la robustesse aux conditions d'utilisation typiques d'un véhicule. De l'image brute (c'est-à-dire pas de correction du *BSF*), jusqu'à l'estimation de la pose, nous avons assemblé et paramétré des méthodes algorithmiques issues de l'imagerie thermique ou visible.

Concernant la réduction du bruit de l'image, rappelons que les tables de correction du *BSF* dépendent de la température du boîtier de la caméra et de la température du plan focal. Nous avons implémenté et évalué des méthodes de l'état de l'art.

Concernant l'estimation de la pose, des maillages 3D du visage, sur lesquels nous avons plaqué de la texture, ont été créés. Ils sont ensuite utilisés pour synthétiser des images dans lesquelles la pose du visage est labellisée. Un premier algorithme, qualifié de « global » et un second, qualifié de « local » ont été implémentés, évalués et comparés entre eux.

Ce prototype (caméra thermique et algorithmes) nous a permis d'évaluer la précision d'estimation de la pose accessible en imagerie thermique. Il nous a également permis de dimensionner les paramètres clefs d'une caméra thermique qui impactent son coût : sensibilité thermique, résolution spatiale.

Mots clefs : surveillance du conducteur, estimation de la pose du visage, imagerie thermique, correction des non-uniformités, approche par points-clefs, approche par apparence globale.

Abstract

This thesis work is about driver head pose estimation thanks to a thermal camera. Head pose estimation covers applications like gaze direction estimation (for driver awareness evaluation), smart airbag (for deployment decision), active head-up display, physiological parameters monitoring (temperature, breathing measurement), etc.

The trend is to use a near infrared illumination (around ~ 900 nm) to operate at night and to be robust to environmental lighting variations. Infrared illumination used is obviously invisible for human eye but to our knowledge no study has been done on possible damage or discomfort on the long term. In this thesis dissertation, we propose alternatively to use uncooled thermal imagery (8-14 µm) to estimate head pose without using any illuminant because thermal cameras detect mainly radiations of hot objects like human faces. Moreover, technological progress in infrared detectors allow a cost reduction, and thermal cameras could become affordable for generalist carmakers like Renault in a few years.

We choose to use a thermal camera based on microbolometer technology. Our first goal is to understand the state of the art methods of spatial noise correction, and algorithms of image analysis to estimate head pose with a highlight on robustness versus use cases in the domain of automotive. From the raw image to the pose estimation, we have put together and tuned methods coming from visible or infrared imaging.

For head pose estimation, virtual 3D shape of the head has been used. Texture has been added to the shape, and the resulting 3D thermal avatar has been used to create synthesized images in which pose is labelled. Two algorithms have been developed to estimate pose, based on the base of synthesized images: the first is called “global”, the second, “local”. Both have been evaluated and compared to each other.

This work allows to evaluate precision of pose estimation which can be reached in thermal imaging. It allows also to define technical specifications like thermal sensitivity and spatial resolution.

Keywords: driver monitoring, head pose estimation, thermal imaging, non-uniformity correction, keypoint approach, global appearance approach.

Sommaire

Introduction

Contexte

Ces travaux de thèse *CIFRE* (Conventions Industrielles de Formation par la Recherche) ont été réalisés au sein de l'équipe *amont ADAS (Advanced Driver Assistance Systems) manœuvre et environnement* sur le site de Guyancourt de l'entreprise Renault.

Que cela soit pour des raisons de sécurité routière, pour des raisons d'interfaçage entre l'homme et la machine ou pour des raisons de délégation de la responsabilité de conduite dans le cadre du véhicule autonome, la surveillance du conducteur au sens large devient un enjeu important pour l'industrie automobile. Pour respecter la contrainte d'acceptabilité par le client, il est nécessaire de privilégier des systèmes sans contact. C'est pourquoi les capteurs optiques et plus particulièrement les caméras sont souvent utilisés. Les conditions d'illumination dans l'habitacle automobile étant très variées et difficiles, un éclairage par Leds autour de 900 nm est adopté dans les systèmes commercialisés et en cours de développement.

Parallélement, les technologies des imageurs thermiques (8 – 14 µm) ont progressé et leur coût a baissé. Un des avantages de l'imagerie thermique par rapport à l'imagerie visible (et proche infrarouge) est la quasi invariance de l'image aux conditions d'illumination. Dans les applications de surveillance des êtres humains, c'est en effet le rayonnement de la peau relatif à sa température, qui est essentiellement numérisé par le capteur. De ce fait, un système de surveillance basé sur une caméra thermique est totalement passif. De plus, une caméra thermique peut donner accès aux évolutions de température (sous certaines conditions techniques de mise en œuvre de l'imageur thermique).

L'une des briques élémentaires des systèmes de surveillance du conducteur par caméra est la fonction d'estimation de la pose du visage. La fonction de pose du visage a pour objectif d'estimer grossièrement la direction de la direction du regard ou d'être une première étape indispensable à une estimation plus fine de la direction du regard qui intégre l'estimation de la position des pupilles dans le globe oculaire. L'estimation de la pose assure un suivi 3D efficace du visage qui s'avère utile dans le cas où les températures de certaines zones telles que le nez ou le front doivent être contrôlées malgré des mouvements amples du visage.

Notons que les applications nécessitant de détecter l'état de l'œil (ouvert ou fermé) ou la position des pupilles dans le globe oculaire ne sont pas adressables par l'imagerie thermique seule car l'œil est relativement homogène en température et difficile à discerner. De plus les verres de lunette ne sont pas transparents aux longueurs d'ondes de la bande thermiques (8 – 14 µm).

Objectif & démarche

L'objectif de ces travaux de thèse est d'étudier la faisabilité de l'estimation de la pose du visage d'un conducteur grâce à un imageur thermique. Les deux axes majeurs qui ont été travaillés sont :

- la mise en œuvre, la maîtrise et l'étalonnage d'une caméra thermique,
- le développement de briques algorithmiques de traitement d'images pour l'estimation de la pose du visage du conducteur.

Pour développer le premier axe, j'ai participé à équiper l'un des laboratoires de Renault avec le matériel nécessaire. Grâce à une collaboration avec le Département d'Optique Théorique et Appliquée (*DOTA*) de l'*ONERA* (Office National d'Etudes et de Recherche Aérospatial) nous avons pu progresser plus rapidement. En effet, nous avons pu effectuer les premières mises en œuvre (prises d'images, étalonnages en chambre climatique) avec le matériel de l'*ONERA*. Nous avons ensuite reproduit et adapté à mon besoin, ce banc d'acquisition d'images très simple. Nous avons par exemple ajouté la possibilité de synchroniser des images thermiques avec une vérité terrain pour l'estimation de la pose du visage.

Les recherches effectuées sur le second axe ont été menées grâce à une collaboration avec le Laboratoire Hubert Curien (*LaHC*) de Saint-Etienne. Nous avons proposé des méthodes d'estimation de la pose : un algorithme dit « global » et un algorithme dit « local », que je présente et évaluons dans ce manuscrit.

Présentation du plan du mémoire

Le manuscrit est composé de six chapitres. Un état des lieux des systèmes d'aide à la conduite est mené au premier chapitre avec un intérêt particulier pour la prise en compte de l'humain dans un véhicule intelligent. Les systèmes de surveillance du conducteur basés sur une caméra seront abordés. L'imagerie thermique, grâce à la réduction des coûts des détecteurs, devient intéressante pour le marché civil. Les caractéristiques de cette filière sont détaillées ainsi que les applications civiles, dont les applications automobiles. Les travaux de recherche permettant d'extraire des paramètres physiologiques humains (augmentation de la pression sanguine, transpiration, rythme respiratoire…) grâce à l'imagerie thermique sont décrits. Enfin, un état de l'art des algorithmes d'estimation de la pose du visage en imagerie visible et en imagerie thermique est présenté. Parmi la littérature nous identifions deux types d'algorithme :

- algorithme « global » : tous les pixels du visage contribuent à l'estimation de la pose,
- algorithme « local » : seulement un certain nombre de pixels du visage, situés à des points clefs, contribuent à l'estimation de la pose.

De cet état de l'art nous remarquons que les algorithmes efficients, qu'ils soient « locaux » ou « globaux » utilisent un maillage 3D pour modéliser la géométrie du visage afin d'augmenter la précision et la rapidité de calcul. J'ai donc choisi d'utiliser également un maillage 3D.

Un détecteur de la filière microbolométrique intégré dans une caméra commerciale a été retenu pour nos recherches. Dans le deuxième chapitre, les principaux paramètres techniques de la caméra thermique (*noise equivalent temperature difference NETD* et bruit spatial fixe résiduel *BSFR*) seront rappelés et mesurées. Une attention particulière est portée sur les méthodes de correction du bruit spatial fixe *BSF*.

Enfin, l'étalonnage radiométrique, au-delà d'être indispensable pour le suivi de température, permet de réduire le temps de calcul des algorithmes de traitement des images. Une expérience simple d'étalonnage radiométrique, utilisant un corps noir dans le champ de vue de la caméra, est présentée.

Le troisième chapitre précise l'utilité du maillage 3D. La « texture thermique » obtenue avec la caméra thermique sera plaquée sur ce maillage. « L'avatar thermique » qui en découle nous permet de créer une base d'images de synthèse personnalisée au niveau de la texture. Les images de synthèse sont des représentations de différentes orientations du visage. D'une part, elles sont labellisées, d'autre part, la troisième dimension, perdue par projection de l'avatar 3D dans le plan image de la caméra virtuelle, peut être retrouvée aisément car le modèle de projection est synthétique et donc connu. Nous avons testé deux maillages 3D, le premier est grossier car il possède une forme ellipsoïdale. Il n'est donc pas adapté à la géométrie 3D du conducteur. Le second est précis et adapté à la géométrie 3D du visage.

Cette base d'images de synthèse sera la donnée de base sur laquelle s'appuient les deux algorithmes d'estimation de la pose développés au quatrième et cinquième chapitre. Le quatrième chapitre détaille une méthode dite « globale » où tous les pixels du visage contribuent à l'estimation de la pose. Cette méthode est basée sur une approche « problèmes inverses ». Le cinquième chapitre décrit une méthode « locale » basée sur l'appariement de points d'intérêt 3D-2D entre l'image à traiter (on parlera d'image réelle) et les images de synthèse. Nous nous appuyons sur les algorithmes du domaine de la vision par ordinateur tels que les algorithmes de détection et de description de points d'intérêt, les algorithmes de résolution du problème *PnP (perspective n points)* et les algorithmes de suppression des erreurs d'appariement (*outliers*).

Enfin nous spécifions, au cours du sixième chapitre, deux caractéristiques techniques de la caméra thermique pour permettre l'estimation de l'orientation du visage : la *NETD* et le format de la matrice de détecteurs (nombre de pixels). Ces deux spécifications impactent le prix d'une caméra thermique. Nous simulerons les dégradations de l'image engendrées par une *NETD* plus importante et un format du détecteur plus petit. La comparaison entre l'algorithme « local » et l'algorithme « global », avant et après avoir appliqué ces dégradations, est menée. Nous proposons notre point de vue sur l'estimation de la pose avec un capteur thermique ayant une résolution et une *NETD* dégradées (par rapport à celui utilisé pour acquérir les images).

Chapitre 1. Surveillance du conducteur par caméra thermique

1.1. La surveillance du conducteur

Mes travaux de thèse s'intègrent dans le cadre général du développement des systèmes d'aide à la conduite communément appelés *ADAS* (de l'anglais *advanced driver assistance systems*). Les *ADAS* agissent sur le comportement du véhicule en analysant les informations acquises à l'aide de capteurs orientés à l'extérieur du véhicule et à l'intérieur de l'habitacle (radar, lidar, caméra, GPS…) et aux capteurs qui décrivent l'état du véhicule (vitesse, accélération…). Le scénario de conduite est analysé et, une alerte, un freinage ou une modification de la trajectoire du véhicule peut être entrepris automatiquement.

La vision du véhicule intelligent par les différents constructeurs premium est intéressante à analyser. Regardons par exemple celle de *BMW*, via le projet *ConnectedDrive* [1]. Le constructeur allemand symbolise le contexte de conduite (*situational context* sur la Figure 1) grâce à un triangle à trois sommets représentant trois pôles d'informations essentielles à recueillir pour concevoir des systèmes intelligents et perçus comme pertinent par le client (on parle d'acceptabilité) : environnement (*environment*), véhicule (*vehicle*) et conducteur *(driver)*. Le pôle « véhicule » récupère, via la méthode de transmission standardisée *CAN* (de l'anglais *Controller Area Network*), les informations telles que l'accélération, la vitesse, la pression des pneus du véhicule. Le pôle « environnement » récupère les informations relatives au trafic, à l'état de la route ou à la signalisation via des capteurs extérieurs (radar, lidar, caméra…). Enfin le pôle « conducteur » concerne les informations telles que la position, la taille, et l'âge du conducteur, ou encore l'orientation de son visage et son état de fatigue.

Figure 1. Projet *ConnectedDrive* de *BMW*. La vision *BMW* du véhicule intelligent intègre des paramètres provenant de 3 pôles représentés par les trois sommets du triangle : l'environnement extérieur (*environment*), l'état du véhicule (*vehicle*) et le comportement du conducteur (*Driver*).

Au delà des *ADAS*, il est intéressant de récolter des informations sur le conducteur pour optimiser d'autres systèmes. Par exemple, les systèmes de confort, tels que la climatisation, ou le préréglage des sièges peuvent être adaptés à la morphologie du conducteur, ou réglés selon ses préférences si il est possible d'identifier le conducteur. Les systèmes *IHM* (interface homme-machine) peuvent également exploiter les informations sur le conducteur. On pense par exemple au contrôle par le geste de certaines commandes, ou

l'adaptation de l'affichage de certaines informations. Les systèmes sécuritaires, comme le système d'airbags, peuvent aussi être optimisés en fonction de la taille et de la posture du conducteur. Enfin avec le développement du véhicule intelligent, la question suivante, relative à la délégation de conduite, se pose : est-ce que le conducteur est en mesure de conduire dans le cas où le véhicule détecte une situation qu'il ne pourra pas gérer ?

Nous avons introduit l'utilité de récolter des informations sur le conducteur pour l'optimisation des *ADAS*, des systèmes de confort, des *IHM* et des systèmes sécuritaires. Nous n'avons pas encore évoqué les autres occupants du véhicule. La connaissance du nombre, de la taille, de la position et même de l'âge des passagers peuvent également permettre d'optimiser les systèmes de confort, des *IHM* et des systèmes sécuritaires. Concernant les systèmes sécuritaires, le déclenchement des airbags pourrait être inhibé ou adapté en puissance selon l'âge du conducteur. La présence d'un siège bébé dos à la route nécessite impérativement d'inhiber le déclenchement des airbags. Le système de climatisation, qui est un système de confort, peut être optimisé grâce à la connaissance du nombre et de la position des occupants. Concernant les *IHM*, le reconnaissance de gestes pour le contrôle des vitres ou encore le contrôle d'un lecteur audio/vidéo est également une forme de surveillance des occupants.

Les nouvelles technologies de capteurs rendent accessibles un plus grand nombre d'informations relatives à l'analyse des êtres humains à l'intérieur de l'habitacle du véhicule. Plus particulièrement, les progrès dans le domaine des capteurs optiques permettent de recueillir ces informations discrètement, c'est-à-dire sans contact, ce qui est indispensable pour des raisons d'acceptabilité. Le véhicule pourrait donc être rendu plus intuitif, plus sécuritaire et plus confortable grâce à l'analyse de occupants.

Mes travaux de thèse portent plus particulièrement sur le conducteur du véhicule. Nous souhaitons récolter des informations permettant d'optimiser le fonctionnement des *ADAS* et également le fonctionnement des *IHM*. En effet, je me suis focalisé sur l'estimation de la pose du visage du conducteur. Celle-ci est intéressante pour plusieurs raisons.

La première raison est que la pose du visage est liée à la direction du regard, qui témoigne du fait que le conducteur est attentif aux dangers potentiels situés autour du véhicule. Les *ADAS* peuvent ainsi être optimisés en anticipant des situations dangereuses que le conducteur n'aurait pas vues. Le terme *human-in-the-loop* est parfois utilisé pour faire référence à ce type d'intégration du comportement humain dans les *ADAS*. De plus, si véhicule détecte que le conducteur a effectivement bien vu le danger potentiel, certains *ADAS*, comme l'alerte de franchissement de ligne, peuvent être inhibés. Ces *ADAS* optimisés ne sont alors plus déconnectés par les clients car il ne sont pas perçu comme gênant et un gain sécuritaire est par conséquent envisageable.

La deuxième raison est que l'estimation de la pose du visage peut être utilisée directement ou indirectement pour détecter la somnolence du conducteur. De manière directe, l'estimation de la pose permet de détecter des hochements de la tête, parfois utilisés pour construire un indicateur de fatigue. De manière indirecte, l'estimation de la pose facilite le suivi de paramètres physiologiques tels que le rythme respiratoire, le rythme cardiaque, la température. Des recherches sont actuellement menées pour créer des

indicateurs de fatigue basés sur le suivi de ces paramètres physiologiques. L'estimation de la pose est une brique essentielle pour rendre le suivi des paramètres physiologiques robuste aux mouvements du visage

La troisième raison est que les systèmes d'affichage tête haute, peuvent également bénéficier de la connaissance direction du regard en adaptant la localisation des affichages sur le pare-brise.

Certains indicateurs relatifs à l'inattention du conducteur sont détaillés dans la section 1.1.1, et ceux relatifs à la somnolence sont détaillés dans la section 1.1.2. Les systèmes d'interfaces homme-machine *IHM* sont indissociables des *ADAS* dans le contexte d'*ADAS human-in-the-loop*. Ainsi, nous mentionnons un système qui exploite la pose du visage dans la section 1.1.3. Nous rappelons très brièvement les enjeux du véhicule autonome du point de vue de la surveillance du conducteur dans la 1.1.4. Enfin dans la section 1.1.5 nous évoquons quelques systèmes commerciaux, basés sur la combinaison de LEDs et d'une caméra, destinés à la surveillance du conducteur. Nous nous focalisons particulièrement sur le matériel et la méthode permettant de détecter la pose du visage et la direction du regard.

1.1.1. Les *ADAS* : vigilance et intention du conducteur

Le système "*driver attention guard*" est un système de maintien de la distance de sécurité avec le véhicule qui précède [2]. Des capteurs à l'extérieur avant du véhicule (Lidar, et caméra) sont capables de détecter un véhicule qui précède ainsi que sa position et sa vitesse différentielle. Ces informations sont fusionnées avec un indicateur de vigilance du conducteur basé sur l'estimation de la direction du regard. Ainsi, lorsque le regard du conducteur quitte la route pendant deux secondes consécutives, le véhicule peut automatiquement adapter sa vitesse si la situation l'exige. Notons que la direction du regard est déduite à partir de l'estimation de la pose du visage du conducteur. Les auteurs signalent que la direction du regard est difficile à estimer de manière robuste. Un indicateur '*eyes on the road / eyes off the road*' est ainsi établi par approximation à partir de la pose du visage. Cette dernière est estimée grâce à deux caméras et un algorithme de traitement d'images [3].

Un autre système a été développé pour prévenir du franchissement de ligne involontaire. Pour réduire le taux de fausse alarme de ce *LDWS* (de l'anglais *lane departure warning system*), l'évaluation de l'intention du conducteur de changer de voie a été integré [4]. Une caméra filme le conducteur et estime la direction du regard ainsi que la direction de la normale au plan du visage. L'intérêt de leurs travaux est de comparer ces deux sources d'informations, c'est-à-dire la direction du regard et la direction de la tête, pour la création d'un indicateur d'intention de changement de voie. La Figure 2 montre les mouvements de la tête (ligne bleue) et les mouvements des yeux (ligne rouge) avant un changement de voie (ligne noire) en fonction du temps. Les images du conducteurs à différents instants sont illustrées sous le graphique. On remarque qu'avant de changer de voie, le conducteur engage d'abord un mouvement de la tête et ensuite un mouvement des yeux. Un délai de 3 s est généralement constaté entre le mouvement de la tête et le début du changement de voie. Lorsque l'on regarde dans l'angle mort, les auteurs expliquent que la variabilité entre les individus, des mouvements des yeux est plus importante que la variabilité des mouvements de la tête. Une des concluions fondamentales de [4] est donc que l'analyse des mouvements de la tête serait plus

pertinente que l'analyse des mouvements des yeux lorsqu'il s'agit d'évaluer l'intention de changer de voie du conducteur.

Remarque : la plupart des constructeurs généralistes désactivent le LDWS lorsque le conducteur active le clignotant avant de changer de voie. Cependant les constructeurs premium estiment que des fausses alarmes dues à une activation tardive des clignotants est gênante pour le client. D'autres méthodes de désactivation du LDWS ont été imaginées. Certaines sont basées sur le calcul du TLC (de l'anglais time to line crossing), c'est-à-dire le temps avant le franchissement de la ligne estimé via l'algorithme d'analyse d'image à partir des images de la route acquises par la caméra frontale. D'autres méthodes ajoutent l'angle du volant. L'estimation des mouvements de la tête est parfois un critère supplémentaire pris en compte par l'algorithme d'apprentissage supervisé dont la fonction est de discriminer l'intention du conducteur de changer de voie ou non [4].

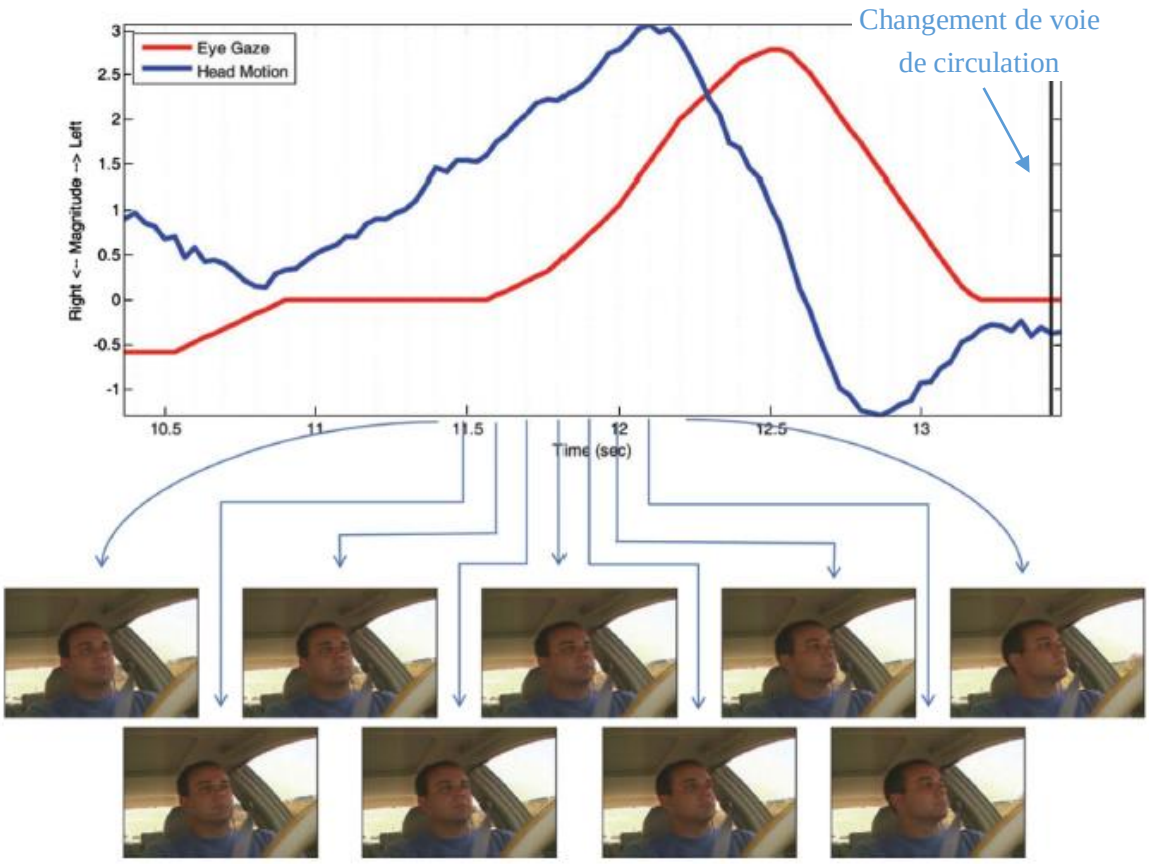


Figure 2. Illustration des mouvements de rotation de la tête en azimut et des changements de direction du regard avant un changement de voie [4]. La courbe bleue représente l'amplitude de la rotation de la tête autour de l'axe vertical. La courbe rouge représente la direction du regard par rapport au repère attaché à la tête. Les photographies représentent les images du conducteur aux instants indiqués par les flèches bleues.

Un indicateur d'attention du conducteur est également utilisé dans un système de freinage d'urgence (*automatic breaking*) et dans un système de maintien latéral du véhicule dans la bande de circulation (*lane keeping)* [5]. Les auteurs considèrent que l'attention du conducteur peut être évaluée à partir de l'estimation

de la direction du regard ou de l'estimation de la détection des mains sur le volant. L'objet de leurs travaux n'étant pas de développer un nouveau système d'estimation de l'attention du conducteur, celle-ci est effectuée en vérifiant que le conducteur utilise ou non un *smartphone* pour écrire un message. Le cœur de leur travaux est de concevoir une loi qui module le déclenchements des systèmes *lane keeping* et *automatic breaking* en fonction de l'attention du conducteur. L'effet d'intégrer le comportement humain dans la chaîne décisionnelle est globalement de réduire le nombre de déclenchement des systèmes *lane keeping* et *automatic breaking* lorsque l'on sait que le conducteur est attentif à sa tâche de conduite.

1.1.2. La somnolence au volant

La somnolence (ou hypovigilance), connue sous le terme anglophone *drowsiness* ou *sleepiness*, est définie comme « le besoin de dormir » [6].

La somnolence au volant est une cause importante d'accidents mortels. D'après l'organisation mondiale de la santé (*OMS*), 1.24 millions de personnes sont mortes sur la route en 2010 et approximativement 6% de ces accidents mortels sont dus à un état de fatigue du conducteur [7]. En Europe, la somnolence au volant est responsable de 20 % des accidents mortels de la circulation [8]. D'après la fondation nationale américaine du sommeil (*US national sleep foundation NSF)*, 28% des conducteurs se sont déjà endormis au volant [9].

La détection robuste et au moment opportun des premiers signes de fatigue est indispensable. Si le système détecte l'endormissement trop tardivement, il est inefficace. Au contraire, si le système est trop précoce dans sa détection de la somnolence, l'acceptabilité sera mauvaise, car les alarmes seront considérées comme de fausses alarmes par les conducteurs qui seront souvent tentés de désactiver le système. Cette détection de la somnolence est une tâche difficile. Nous présentons ci-dessous quelques exemples de systèmes proposés par des recherches académiques ou déjà commercialisés.

L'analyse de l'angle du volant est utilisé pour évaluer le niveau de fatigue du conducteur. Citons par exemple les systèmes de Mercedes-Benz [10] et de Volkswagen [11]. Lorsqu'il s'endort, le conducteur à tendance à ne plus adapter la trajectoire du véhicule à la route, jusqu'à une correction abrupte. Les systèmes de Mercedes-Benz et Volkswagen analysent la vitesse de l'angle du volant en fonction du temps. Lorsque des motifs spécifiques sont répétés au delà d'un certain seuil, une alerte peut être envisagée.

D'autres constructeurs ont opté pour des systèmes de surveillance basés sur une caméra de surveillance du conducteur. Dès 2006 le groupe Toyota (via Lexus) propose un système nommé *DMS* pour *driver monitoring system* qui identifie si le conducteur regarde la route ou non. Combiné à un détecteur d'obstacle, des mesures d'alerte, de freinage d'urgence ou de préparation à l'impact peuvent être activées. Finalement, ce système estime le niveau d'attention du conducteur, mais il n'est pas spécifique à la détection de l'endormissement, c'est-à-dire à la survenue de somnolence. En 2008 Toyota intègre un détecteur de fermeture de paupières pour accéder à des informations plus spécifiques à l'endormissement. C'est donc l'estimation de l'état des yeux (ouverts ou fermés) qui permet de construire des indicateurs de somnolence.

Un autre exemple de système d'alerte de fatigue du conducteur repose sur deux indicateurs [12]. Le premier est basé sur le taux et le temps de fermeture des yeux, comme dans le cas de Toyota. Le second est basé sur l'analyse des hochements de tête (cf. Figure 3). Lorsque la fréquence des hochements dépasse un seuil, une alerte peut être donnée au conducteur.

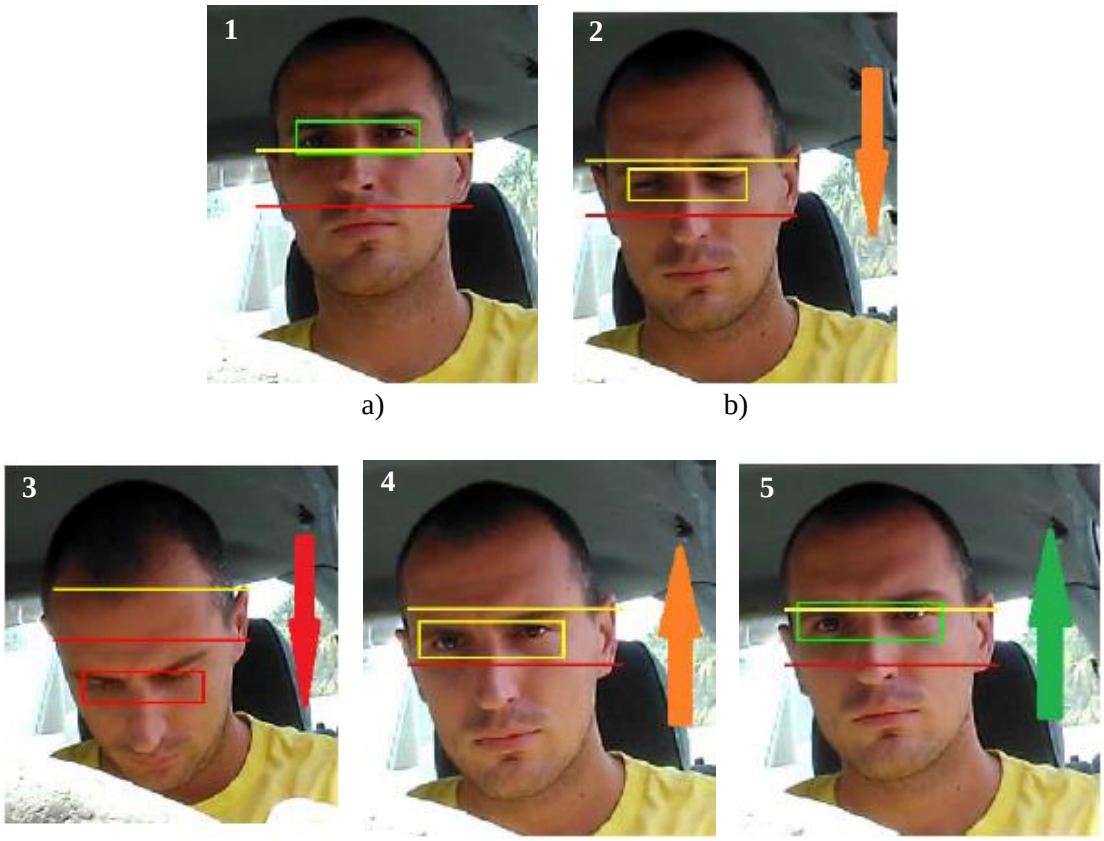
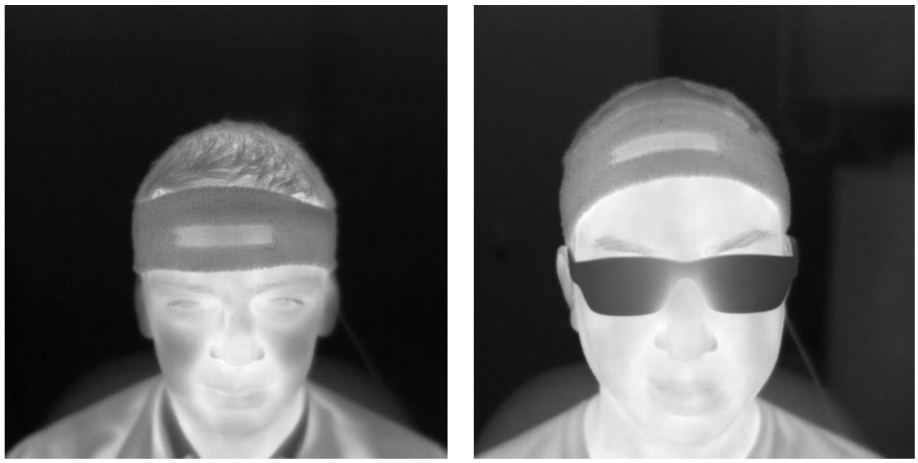


Figure 3. Hochement de la tête typique d'une phase d'endormissement du conducteur.

L'évaluation du niveau d'endormissement grâce à l'estimation de la pose du visage est donc possible, mais aucun système n'utilise uniquement cette information car la robustesse de la détection serait probablement insuffisante. Nous insistons sur ce point car nous avons fait le choix d'utiliser une caméra thermique dans ces travaux de thèse. Or, les yeux sont très peu discernables en imagerie thermique comme on peut le voir sur la Figure 4 a). De plus, les verres de lunettes qu'ils soient solaires ou pour la correction de la vue, absorbe presque totalement le rayonnement thermique comme on peut le voir sur la Figure 4 b). Cependant, une caméra thermique permet d'obtenir des informations sur les variations locales de la température du visage. Cela ouvre des perspectives pour la surveillance de paramètres physiologiques, qui seront détaillées dans la section 1.4, et qui peuvent être reliés au niveau de fatigue du conducteur.



a) b)

Figure 4. Images thermiques d'ingénieurs de Renault. a) Individu sans lunette. b) Indiviu portant des lunettes de vue. Notons que le verre absorbe totalement le rayonnement thermique.

Il est communément admis qu'une augmentation de la fatigue est accompagnée d'une diminution de la fréquence respiratoire [13]. Récemment l'équipementier automobile espagnol *FICOSA* en partenariat avec l'institut biomécanique de valence (*IBV* pour *Institute of Biomechanics of Valencia*) a implémenté un système d'estimation du rythme respiratoire du conducteur [14]. Leur méthode, testée uniquement en laboratoire, est basée sur la mesure des mouvements de la poitrine grâce à une *Kinect*. Un aspect important de leur travail est de rendre le système robuste aux variations d'illuminations dans l'habitacle du véhicule.

L'analyse des pulsations cardiaques est un autre paramètre qui évolue en cas de fatigue. La variabilité de la fréquence cardiaque *HRV* (pour *heart rate variability*) est un indicateur parfois utilisé pour détecter la fatigue du conducteur [15,16].

La direction du regard, on l'a vu, est liée au niveau d'attention du conducteur à sa tâche de conduite. Parallèlement à cela, les véhicules capables aujourd'hui de discerner certains dangers, donnent de plus en plus d'informations visuelles pour en informer le conducteur. Une réflexion sur la conception des *IHM* est menée chez les différents constructeurs pour que cette transmission d'informations supplémentaires par le canal visuel ne soit pas une cause de baisse ou de perte d'attention du conducteur vis-à-vis de sa tâche de conduite. Cette problématique montre qu'*ADAS* et *IHM* sont intrinsèquement liés.

1.1.3. Interface homme machine

Il existe de nombreuses manières d'interagir avec le conducteur. On peut citer par exemple une interface haptique (vibration du volant ou de la pédale de frein par exemple), ou encore une interface audio. Cependant, l'interaction visuelle est souvent privilégiée dans un habitacle automobile car elle permet de transmettre une quantité d'informations très importante. Le véhicule intelligent doit informer de manière pertinente en distrayant au minimum le conducteur. Certains constructeurs, en s'inspirant de l'aviation, ont développé des afficheurs tête haute également appelés *HUD* (de l'anglais *head-up display*). Présent chez

les constructeurs premium, les *HUDs* proposent d'afficher les informations sur le pare-brise du véhicule à une localisation fixe, on parle alors de *HUD* statique [17].

Les prochaines générations de *HUD* intègreront des informations sur le conducteur. Il pourrait notamment être question d'adapter dynamiquement l'affichage (forme, couleur, taille) en fonction du contexte ou d'adapter la localisation des informations en fonction de la direction du regard. Concernant le deuxième point, il a été montré que la localisation idéale pour afficher des informations autres que des alertes de danger se situe à 5° à droite de la direction principale du regard [18].

La référence [19] exploite un système d'estimation de la pose de la tête pour adapter la localisation et l'intensité des informations d'un *HUD*. Ces travaux portent sur l'alerte de dépassement de vitesse. La rapidité d'adaptation de la vitesse du véhicule à la vitesse limite autorisée ainsi qu'une distraction moins importante du conducteur vis-à-vis de sa tâche de conduite (mesurée objectivement par le rapport du temps passé à regarder la route sur le temps passé à regarder l'information de vitesse) sont les avantages essentiels des *HUD* actifs par rapport à un *HDD* (*Head-down display*).

1.1.4. Le véhicule autonome

Le développement du véhicule tout autonome semble inéluctable [20]. Cependant, un déploiement à l'échelle mondiale d'un véhicule avec une autonomie importante serait à ce jour prématuré. La stratégie adoptée par la plupart des constructeurs automobile est une augmentation progressive de l'autonomie du véhicule. La personne assise à la place du conducteur sera donc invitée périodiquement à contrôler manuellement le véhicule, soit parce qu'elle le souhaite, soit parce que le véhicule est inapte au mode autonome dans la situation en cours ou à venir [21]. Dans les deux cas, avant de déléguer au client la responsabilité de la conduite, il est nécessaire de s'assurer de sa capacité à conduire. En particulier, s'il n'est pas attentif à la scène routière et/ou s'il n'est positionné correctement par rapport aux commandes du véhicule, des systèmes d'avertissement doivent le prévenir. Cela sous-entend que le véhicule soit équipé de systèmes permettant de surveiller certains paramètres liés à l'état du conducteur.

Plusieurs systèmes sont envisagés tels que les systèmes de détection des mains sur le volant (*HOD* de l'anglais *hands on/off detection*), de détection des pieds sur les pédales ou encore de l'estimation de la direction du regard. Les enjeux sont, comme bien souvent dans le domaine automobile, de s'assurer de :

- la robustesse et de la fiabilité de ces systèmes pour des raisons de sécurité évidentes
- l'acceptabilité de ces systèmes par le client.

Dans ce manuscrit nous nous focaliserons sur les systèmes d'estimation de la direction du regard. Dans cette perspective, une technique possible consiste à utiliser une caméra qui filme le visage du conducteur. Cette approche est souvent explorée par les constructeurs automobiles car elle est sans-contact et, par conséquent, ne génère aucune gêne pour le conducteur.

1.1.5. Quelques solutions commerciales d'estimation de l'état du conducteur

Citons, de manière non-exhaustive, un certain nombre d'entreprises qui proposent des logiciels de suivi (*tracking*) et d'analyse du visage : *SeeingMachine*, *VisageTechnologies*, *SmartMeUp*. Le point commun des logiciels des entreprises citées est le recours à des algorithmes d'apprentissage supervisé pour la détection de points caractéristiques du visage tels que les coins du nez, les contours du visage et de la bouche… Ces points peuvent également être appelés points d'ancrage, points spécifiques, amers …

Voici une liste non-exhaustive des équipementiers automobiles qui développent des solutions d'estimation de la direction regard du conducteur par caméra : *AISIN*, *Omron*, *Autoliv*, *Valeo*… Citons toujours de manière non-exhaustive, quelques sociétés, qui se sont spécialisées dans le développement de systèmes complets (caméra, illuminateurs et logiciel) pour la surveillance du conducteur par caméra : *SmartEye* (cf. Figure 5 et référence [22]) et *InnovPlus* (cf. Figure 6 et référence [23]). Cette dernière est une start-up française située sur le plateau de Saclay.

Figure 5. Système *AntiSleep* proposé par la société *SmartEye* en 2014 [22]. Une caméra sensible dans la bande visible et *NIR* est associée à deux groupes de Leds placés à gauche et à droite de la caméra.

Figure 6. Système de surveillance du conducteur par caméra *Toucango* de la société *InnovPlus* [23]. Ce système se fixe sur le pare-brise grâce à une ventouse à la manière d'un GPS.

Les systèmes développés, ou en cours de développement, fonctionnent généralement grâce à une caméra sensible aux longueurs d'ondes visibles et *NIR* (*near infrared*, entre 0.7 et 1.4 µm) car des imageurs à un coût accessible sont déjà disponibles chez les fabricants. Le secteur de la téléphonie mobile a effectivement joué un rôle important dans la baisse des coûts des capteurs visibles et *NIR*. Afin que le système fonctionne la nuit et soit robuste aux conditions d'illuminations difficiles de jour, des LEDs émettant autour de 0.9 µm génèrent un éclairement constant du visage du conducteur. On peut voir ces LEDs sur la Figure 5 et sur la Figure 6. L'effet des LEDs est illustré sur la Figure 7, qui est issue de la référence [24]. Sur la première ligne, on peut voir les images d'un visage éclairé de manière inhomogène par l'environnement. Ces images sont acquises avec seulement une caméra, c'est-à-dire sans utiliser de LEDs. Sur la seconde ligne de cette figure, les images sont acquises alors que les LEDs *NIR* éclairent le visage, sachant que les conditions d'éclairage de l'environnement sont identiques que sur la première ligne.

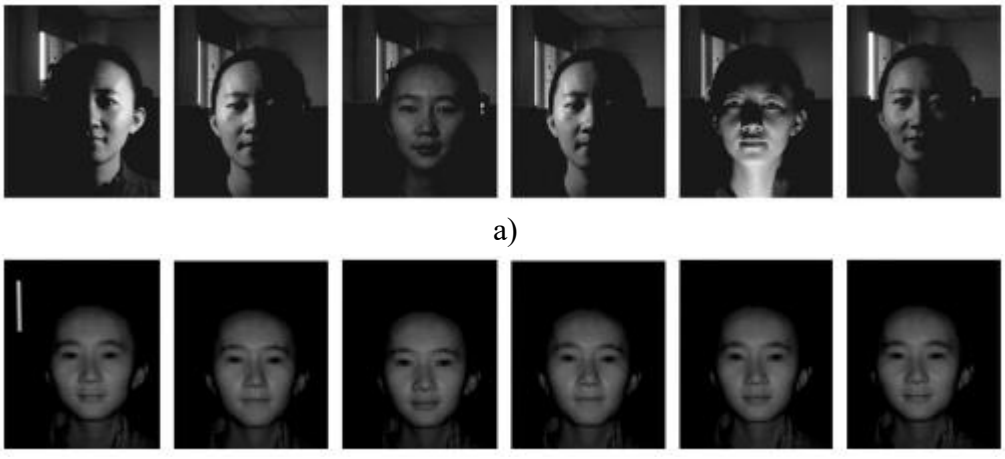


Figure 7. Illustration de l'effet des Leds *NIRs* lorsque les conditions d'illumination naturelle sont difficiles : a) les Leds *NIRs* ne sont pas utilisées, b) les Leds *NIRs* sont utilisées. Les conditions naturelles d'illuminations sont les mêmes pour un couple d'images d'une même colonne. Ces images sont issues de la référence [24].

Cette nouvelle génération de produit dit *DMS* (*driver monitoring system*) proposent généralement plusieurs niveaux d'informations. Ils sont capables :

- d'estimer la pose de la tête,
- d'estimer la direction du regard,
- d'estimer l'état des yeux (fermés/ouverts).

L'estimation de la pose de la tête est la plus facile à réaliser car c'est un objet de taille relativemement importante lorsque le champ de vue de la caméra est centré et resseré sur lui. L'estimation de la direction du regard est plus difficile car une résolution spatiale plus importante est nécessaire. Généralement on identifie la position de la pupille par rapport aux bords de l'œil, ce qui implique d'être capable de détecter les coins des yeux [25].

La position de la pupille dans le globe occulaire donne accès à la direction du regard par rapport à la normale au visage. Ainsi, la direction réellement ciblée par le conducteur est, au premier ordre, donnée par la direction de la tête. Au second ordre, la position des pupilles affine l'estimation. Cette combinaison est illustrée par la Figure 8 en imagerie visible. Notons que le principe est similaire en imagerie *NIR*.

L'estimation de la direction du regard souffre de quelques limitations. En effet, lors de rotations amples de la tête, les yeux ne sont plus visibles par la caméra. Ainsi, dans ces situations, l'estimation du regard ne peut être approximée qu'avec l'estimation de la pose de la tête [4]. Comme nous l'avons déjà évoqué, dans les situations de changement de direction important du regard, le mouvement de la tête est un excellent indicateur pour évaluer les intentions du conducteur. Il est même aussi bon (voir meilleur dans certaines

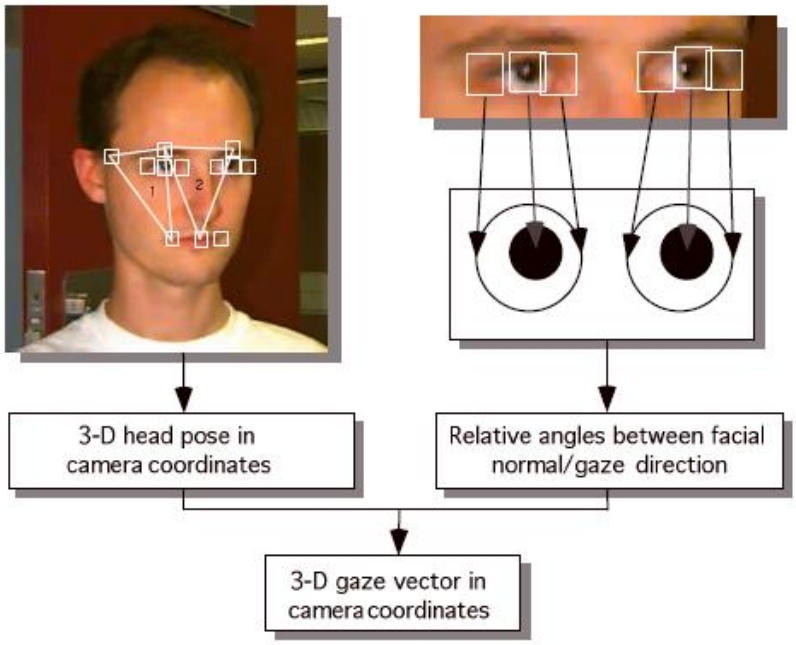


Figure 8. Exemple typique de méthode d'estimation de la direction du regard [25]. La pose du visage est dans un premier temps évaluée, puis l'angle relatif entre la normale au visage et la direction des yeux vient moduler l'estimation de la pose du visage pour donner l'estimation finale de la direction du regard.

applications) qu'un indicateur qui intègre le suivi (*tracking*) des yeux comme le confirme l'étude [26] menée par l'agence fédérale américaine *NHSTA* (*national highway traffic safety administration*).

L'estimation du regard implique également d'être capable de détecer les coins des yeux, qui d'une population à une autre peut différer, et nécessite donc un apprentissage supervisé important, qui augmente le coût de développement final du système.

La direction du regard reste cependant essentielle pour détecter un certain type de distraction. En effet, lors de l'utilisation d'un *smartphone* ou de la tablette centrale du véhicule (c'est-à-dire l'*IHM* principale de nombreux véhicules actuels), nous avons naturellement tendance à privilégier une faible rotation du visage et des mouvements oculaires de grandes amplitudes. Donc, dans le cas de la surveillance de l'inattention l'estimation de l'orientation de la tête ne semble pas suffisante. Cependant, même si les progrès en traitement d'images sont importants, il est difficile de garantir la fiabilité d'un système dédié à l'estimation de la direction du regard.

Concernant l'estimation de l'état ouvert ou fermé de l'œil, des algorithmes basés sur de l'apprentissage supervisé sont proposés. Comme pour l'estimation de la position des pupilles, l'apprentissage supervisé peut être vu comme une difficulté.

L'imagerie visible/*NIR*, pour la surveillance du conducteur doit être impérativement combinée à des LEDs *NIR* afin d'assuer un fonctionnement la nuit et afin d'assurer la robustesse aux conditions d'illuminations difficiles le jour. Bien que l'imagerie thermique ne permette pas de récolter des informations sur les yeux, nous allons voir dans la section suivante, qu'elle possède l'avantage de fonctionner de manière totalement passive, c'est-à-dire sans LED.

1.2. L'imagerie thermique pour la surveillance de l'être humain

Les LEDs *NIR* utilisées pour éclairer le visage du conducteur (cf. section 1.1.5) respectent normalement les normes de sécurité pour protéger les yeux des conducteurs. Cependant, à notre connaissance, aucune étude évaluant les risques d'une utilisation sur le long terme n'existe. L'imagerie thermique fonctionne sans utiliser de LEDs car elle détecte le rayonnement propre de la scène. Elle permet donc de résoudre sans équivoque cette problématique de sécurité oculaire posée par l'utilisation des LEDs *NIR*. C'est en partie pour résoudre ces problèmes de sécurité oculaire que nous avons souhaiter étudier la surveillance du conducteur par imagerie thermique.

1.2.1. Bande spectrale infrarouge thermique (*LWIR*, 8-14 µm)

La bande spectrale infrarouge s'étend de 0.7 µm à 1 mm. Ce rayonnement est invisible pour notre œil. L'atmosphère absorbe et transmet certaines gammes de longueurs d'onde (cf. Figure 9). On distingue quatre bandes pour lesquelles la transmittance de l'atmosphère est élevée :

- Le *NIR* (*near infrared*) est compris entre 0.7 et 1.4 µm
- Le *SWIR* (*short wavelength infrared*) est compris entre 1.4 et 3 µm
- Le *MWIR* (*middle infrared wavelength*) est compris entre 3 et 5 µm

- Le *LWIR* (*long wavelength infrared*) est compris entre 8 et 14 µm
- Le *FIR* (*far infrared*) est compris entre 15 µm et 1 mm

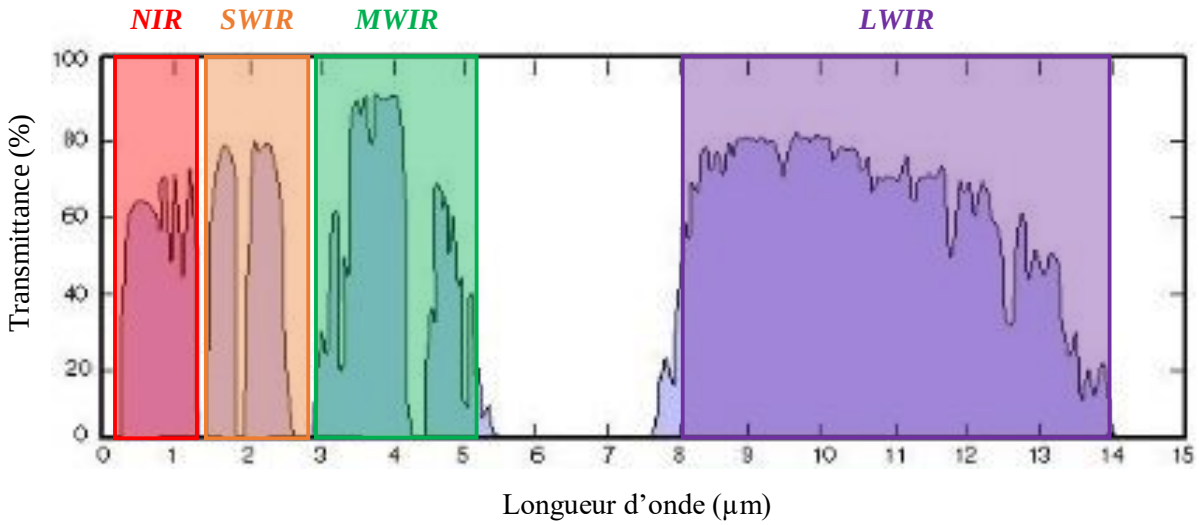


Figure 9. Répartition des bandes de transmission de l'infrarouge dans l'atmosphère.

Des technologies de détecteurs infrarouges ont été développées dans ces différentes fenêtres de transmission. Lors de la conception d'un système, le choix de la longueur d'onde doit être réalisé pour faciliter la discrimination de l'objet par rapport au fond. Dans notre application la peau humaine est l'objet d'intérêt. Celle-ci possède une température proche de 30°C. Grâce à la théorie du corps noir [27, 28], il est possible d'approximer le flux spectrique rayonné par la peau humaine. Tout corps à une température différente du zéro absolu (c'est-à-dire différente de 0 K) rayonne du fait de l'agitation thermique. La luminance est une grandeur radiométrique utilisée pour quantifier la quantité d'énergie rayonnée par un point source, par unité de surface et par unité d'angle solide. Un corps noir est un objet dont l'émissivité vaut 1 et qui rayonne de manière isotrope. La loi de Planck exprime la luminance spectrique L_λ en $W.m^{-2}sr^{-1}.m^{-1}$ d'un corps noir idéal en fonction de sa température T (cf. Annexe A). La Figure 10 représente la luminance spectrique d'un corps noir à 280 K, 300 K et 320 K. On remarque que les fenêtres *MWIR* et *LWIR* semblent les mieux adaptées pour détecter des corps aux alentours de 30°C. Actuellement, les détecteurs *MWIR* sont basés sur une technologie de détection photonique (ou quantique) refroidie (cf. section **Erreur ! Source du renvoi introuvable.**). A ce jour, le coût de cette technologie reste prohibitif p our l'industrie automobile. Au contraire, dans le *LWIR* une gamme de détecteurs appelée thermique (cf. section 1.2.2) n'a pas besoin d'être refroidie. Cela participe au fait que leur prix semble plus abordable pour l'industrie automobile. Ces raisons font que le *LWIR* est souvent préféré pour la détection de corps chaud dans le marché civil (cf. section 1.3.2).

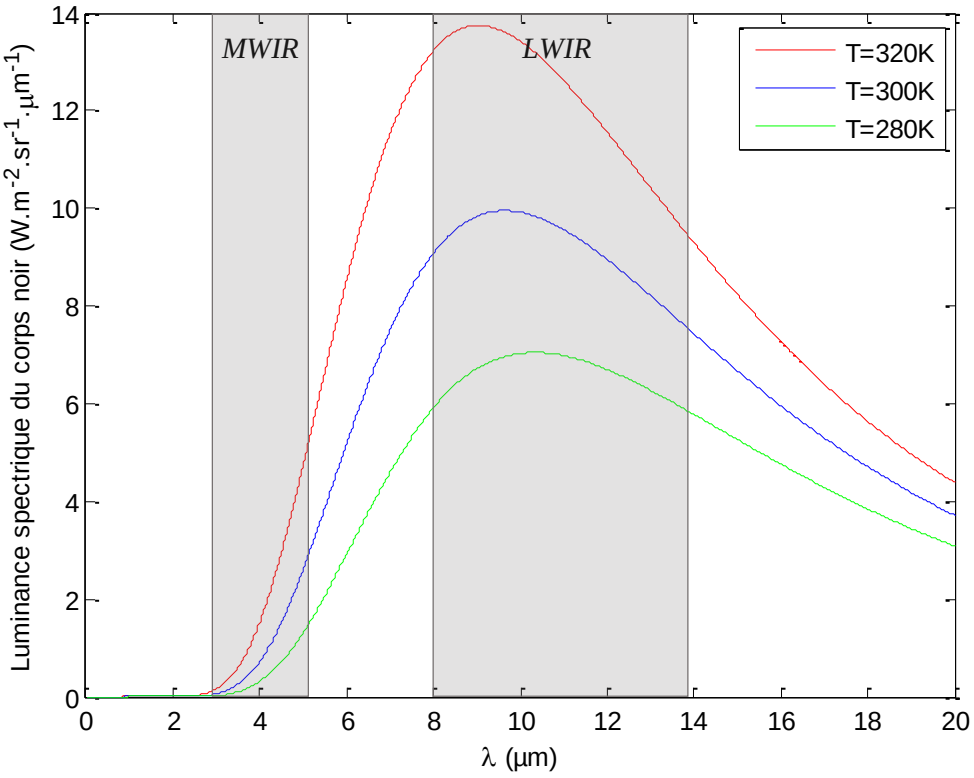


Figure 10. Luminance spectrique d'un corps noir à différentes températures.

Enfin terminons par l'apport de quelques précisions concernant l'émissivité. Celle-ci dicte la capacité d'un corps à rayonner. Elle est définie comme le rapport entre la luminance spectrique d'un corps noir idéal L_λ^{CN} et celle d'un corps réel :

$$\varepsilon = \frac{L_\lambda}{L_\lambda^{CN}} \tag{1.1}$$

Ainsi la quantité d'énergie rayonnée par un objet réel, par unité de surface d'angle solide est :

$$L_\lambda = \varepsilon . L_\lambda^{CN} \tag{1.2}$$

Remarque : une modélisation plus précise de l'émissivité communément utilisée prend en compte la dépendance angulaire. Nous donnons volontairement ici une explication simple de l'émissivité.

L'émissivité de la peau est proche de 1, celle de cheveux, des sourcils et de la barbe est très légèrement inférieure. Une étude portant sur l'émissivité de la peau des truies a montré que la présence de poils diminue l'émissivité de l'ordre de 0.02 [29]. Lorsque que l'on observe l'image thermique d'un visage, un contraste entre ces éléments et la peau apparait. Il est dû à une différence de température et d'émissivité.

L'utilisation de l'imagerie *LWIR* garantit un fonctionnement passif (c'est-à-dire sans l'utilisation d'un illuminant artificiel). En effet, la peau ne transmet pas (ou très peu) et ne réfléchit pas (ou très peu) le rayonnement thermique dans le *LWIR*. Elle est donc un bon absorbant d'après le principe de conservation

de l'énergie. D'après la loi de Kirchhoff qui exprime le fait qu'un bon absorbant est un bon émetteur, la peau possède une bonne capacité à rayonner. De plus, son émissivité est proche de 1, ce qui fait d'elle un bon corps noir. Ainsi, lorsque l'on observe dans le *LWIR* la peau d'un être humain, on observe son rayonnement propre qui dépend de sa température. L'image de la peau est donc :

- indépendante des conditions d'illuminations,
- dépendante des paramètres physiologiques qui modulent la température d'un individu.

Pour nous rendre compte visuellement de l'invariance de l'image aux conditions d'illumination, nous avons mis côtes à côte des images thermiques et visibles illustrées à la Figure 11, tirées de la base publique réalisée par H. Chang *& al* [31]. Les images thermiques ont été acquises avec une caméra *LWIR* de marque *Raytheon*, modèle *Palm-IR-Pro* au format 320×240 pixels présentant une résolution thermique de 100 mK. Les conditions d'illuminations naturelles sont identiques sur les deux images d'une ligne de la Figure 11.

La couleur de peau n'impacte pas suffisamment l'émissivité pour que cela soit perceptible par un imageur thermique fonctionnant dans le *LWIR* (cf. base H. Chang *& al* [31]). C'est-à-dire que le contraste entre le fond et le visage reste toujours très important (lorsque le fond est à température ambiante).

Notons tout de même quelques limitations à la reproductibilité des images thermiques d'un visage. Dans la référence [32] les auteurs ont constaté dans un contexte automobile que lorsque la fenêtre est ouverte, si la température de l'air est très différente de celle de la peau, la température de cette dernière peut être modifiée localement. De plus, sans avoir besoin de se référer à la bibliographie scientifique, nous pouvons anticiper certains problèmes en cas de sudation importante sur le visage car l'eau possède une émissivité (entre 0.95-0.963) inférieure à celle de la peau (souvent considérée à 0.98). Ces variations d'apparence thermique du visage sont cependant plus lentes que les variations soudaines d'illuminations ce qui laisse la possibilité aux algorithmes de se mettre à jour.

Figure 11. Illustration de l'invariance de l'imagerie thermique à l'illumination naturelle. Ces images proviennent de la base de données publique réalisée par H. Chang & *al* [31]. Les conditions d'illuminations naturelles sont identiques sur les deux images d'une ligne

1.2.2. Les détecteurs thermiques (ou détecteurs non-refroidis)

Un détecteur thermique absorbe le flux infrarouge émis par la scène, ce qui provoque son échauffement, et une modification d'un paramètre physique. Un circuit de lecture judicieusement conçu permet de détecter les variations du paramètre physique. Ce type de détecteur fonctionne à température ambiante, ce qui le rend beaucoup plus accessible pour le marché civil car il est moins lourd, moins encombrant et surtout moins cher qu'un détecteur infrarouge quantique qui doit être refroidi à des températures inférieures à 110 K et qui nécessite une machine cryogénique. La sensibilité des détecteurs thermiques est cependant inférieure à celle des détecteurs quantiques. Mais pour la plupart des fonctionnalités recherchées dans les applications civiles, les besoins en termes de résolution thermique ne sont pas aussi élevés que ceux du marché militaire. Ce sera plutôt, le coût qui sera un critère primordial.

On distingue quatre filières de détecteurs thermiques : la filière thermopile, la filière pyrométrique, la filière des microbolomètres et la filière des thermodiodes. Le rapport entre le nombre de pixels par matrice et le prix est illustré pour les différentes filières sur la Figure 12 [33].

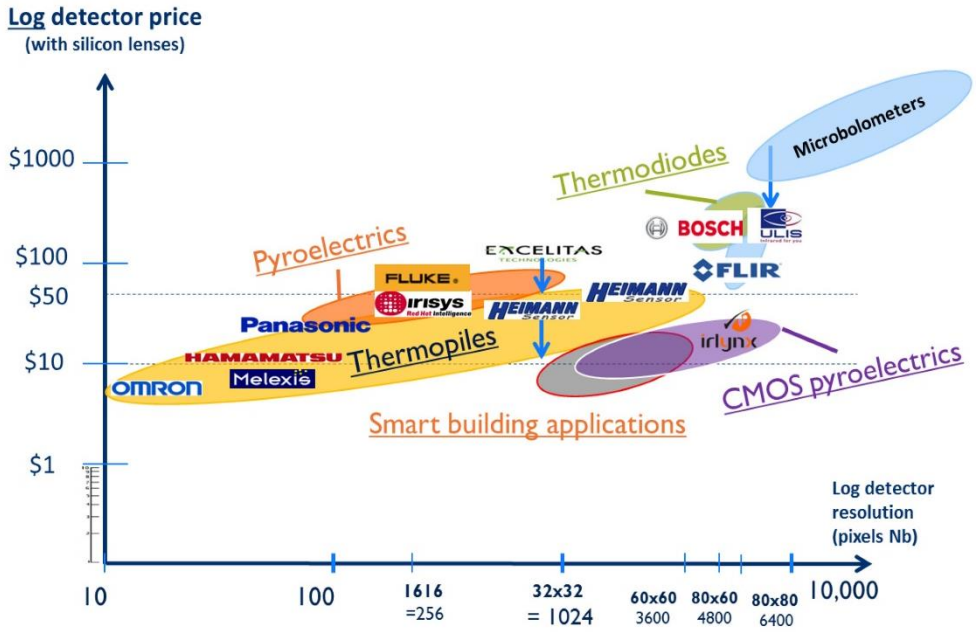


Figure 12. Illustration des détecteurs thermiques à bas coût disponibles sur le marché [33]. Le coût des détecteurs en fonction du nombre de pixels est représenté sur ce graphique. Notons que le pas pixel ainsi que la sensibilité thermique des détecteurs ne sont pas représentés.

Une catégorie de détecteur thermique repose sur l'effet pyroélectrique. Sous l'effet de la chaleur, la structure cristalline du matériau se dilate et une charge surfacique se créé. Lors du retour à l'équilibre, un courant apparaît spontanément si le matériau transducteur est intégré dans un circuit de lecture. Ces détecteurs sont donc sensibles à des variations temporelles du rayonnement de la scène. Ainsi, pour observer une scène fixe, un obturateur mécanique peut être utilisé pour créer une variation artificielle du rayonnement de la scène. L'effet pyroélectrique disparait lorsque le détecteur dépasse la température de Curie. La start-up française *Irlynx* tente de mettre sur le marché une matrice pyroélectrique compatible avec la technologie

CMOS (de l'anglais *complementary metal oxide semi-conductor*) avec un rapport prix/nombre de pixels avantageux par rapport aux détecteurs de la filière microbolomètrique.

La deuxième filière est basée sur la technologie thermopile. Un pixel thermopile est composé de plusieurs thermocouples (de 6 à 16 d'après le brevet [34]). Un thermocouple est composé de deux matériaux de nature différente qu'on appelle 'soudure chaude' et 'soudure froide'. La 'soudure chaude' est placée dans le milieu dont on cherche à mesurer la température. La 'soudure froide' est isolée thermiquement de la 'soudure chaude'. Grâce à l'effet Seebeck, la différence de température entre les deux soudures fait varier la tension électrique. Un pixel thermopile comporte une surface qui absorbe le rayonnement thermique de la scène et qui est accolée aux 'soudures chaudes' (des 6 à 16 thermocouples). La température est mesurée relativement par rapport aux soudures froides placées dans le boîtier électronique du système. Généralement les matrices possèdent peu de pixels et ont une moins bonne résolution thermique que les microbolomètres. A notre connaissance les matrices de capteurs proposées par la société *Heimann sensor* sont celles qui comptent le plus en plus de pixels dans cette filière. Leur format est 82×62 pixels au pas de 100 µm à 9 Hz avec une résolution thermique de 115 mK.

La troisième filière est basée sur la technologie thermodiode. Lorsque qu'une diode est passante (c'est-à-dire polarisée en tension), un courant peut la traverser. Celui-ci dépend de la température du composant. Un pixel thermodiode est composé d'une ou plusieurs diodes qui sont accolées à une couche qui absorbe le rayonnement thermique de la scène. *Bosch* propose le produit *SMO130* au format 82×62 pixels au pas de 100 µm avec une résolution thermique de 200 mK @f/1 et @9Hz. Le *SMO130* fonctionne dans une plage de température environnementale de -20°C à +65°C. L'intention de *Bosch* est d'atteindre le prix agressif de 10 € pour un système qui comprend une matrice de 100×50 pixels avec l'optique [35]. Ce prix a été défini lors du projet européen *ADOSE* (*reliable Application-specific Detection of road users with vehicle On-board SEnsors)* [36]. Ce type de caméra serait destiné à la détection de corps chauds à l'extérieur du véhicule. Dans le projet *ADOSE* la caméra thermique compléterait des systèmes de détection basés sur la caméra visible/*NIR* qui peuvent être mis en défaut dans certains cas d'usages, comme une météo difficile.

Enfin, la quatrième filière est basée sur la technologie microbolométrique. Initialement réservée au domaine militaire, elle est coûteuse mais ses performances en termes de résolution thermique et spatiale sont les meilleures parmi toutes les filières de l'imagerie infrarouge non refoidie. Un matériau bolomètre possède une résistance électrique qui dépend de la température. Cette dépendance suit la loi empirique d'Arrhenius. Le matériau transducteur est accolé à une surface qui absorbe le rayonnement *LWIR* de la scène. Afin que les pixels soient sensibles au rayonnement de la scène, il est nécessaire de les isoler thermiquement. Cela se traduit par une mise sous vide d'un espace qui contient la matrice de pixels et qui est délimité par un hublot illustré sur la Figure 13.

Afin de réduire le coût, la société française *Ulis* a développé à ce sujet une maîtrise du vide à l'échelle du pixel (cf. projet *MIRTIC mirco retina thermal infrared circuit* détaillé à la section 1.3.2). Aini, le hublot est supprimé.

Une des caractéristiques de la technologie microbolométrique, est qu'il est nécessaire de polariser chaque pixel, ce qui engendre une partie du bruit spatial fixe *BSF* du détecteur. Nous abordons en détail les

méthodes de correction *BSF* au Chapitre 2, section 2.5. La particularité du *BSF*, en imagerie non-refroidie, est qu'il est dépendant de l'état thermique du bloque optique, du bloque de détection et du bloque électronique. Les différents bloques qui composent une caméra thermique sont illustrés sur Figure 13. Afin de simplifier la correction du *BSF,* un refroidissement thermoélectrique *TEC* (*thermoelectric cooling*) par module Peltier, illustré sur la Figure 13, permet stabiliser la température du bloque de détection. De plus, pour un obturateur mécanique, appelé *shutter*, peut être déclanché pour positionner un corps uniforme devant la matrice de pixels. Ainsi, estimant les disparités spatiales des réponses des pixels lorsque le *shutter* est déclenché, le correction du *BSF* peut être mise à jour.

Pour des raisons de coût les méthodes de correction du *BSF* ont évoluées au cours du temps. Dans un premier temps, le module Peltier a été supprimé et seul le *shutter* a été conservé. Ainsi, en fonctionnement *TEC-Less*, c'est-à-dire sans le module Peltier, la correction du *BSF* est dès le démarrage de la caméra en déclenchant le *shutter*. Ensuite, lorsque la température de la caméra évolue au-delà d'un seuil, le *shutter* est déclenché une nouvelle fois, et la correction du *BSF* est mise à jour. Ce type de correction nécessite de déclencher souvent le *shutter*, ce qui peut être gênant dans certaines applications car la caméra n'est pas opérationnelle pendant cette phase.

Depuis peu, des algorithmes de traitement d'images combinés à un étalonnage permettent de d'envisger une suprression du *shutter*.

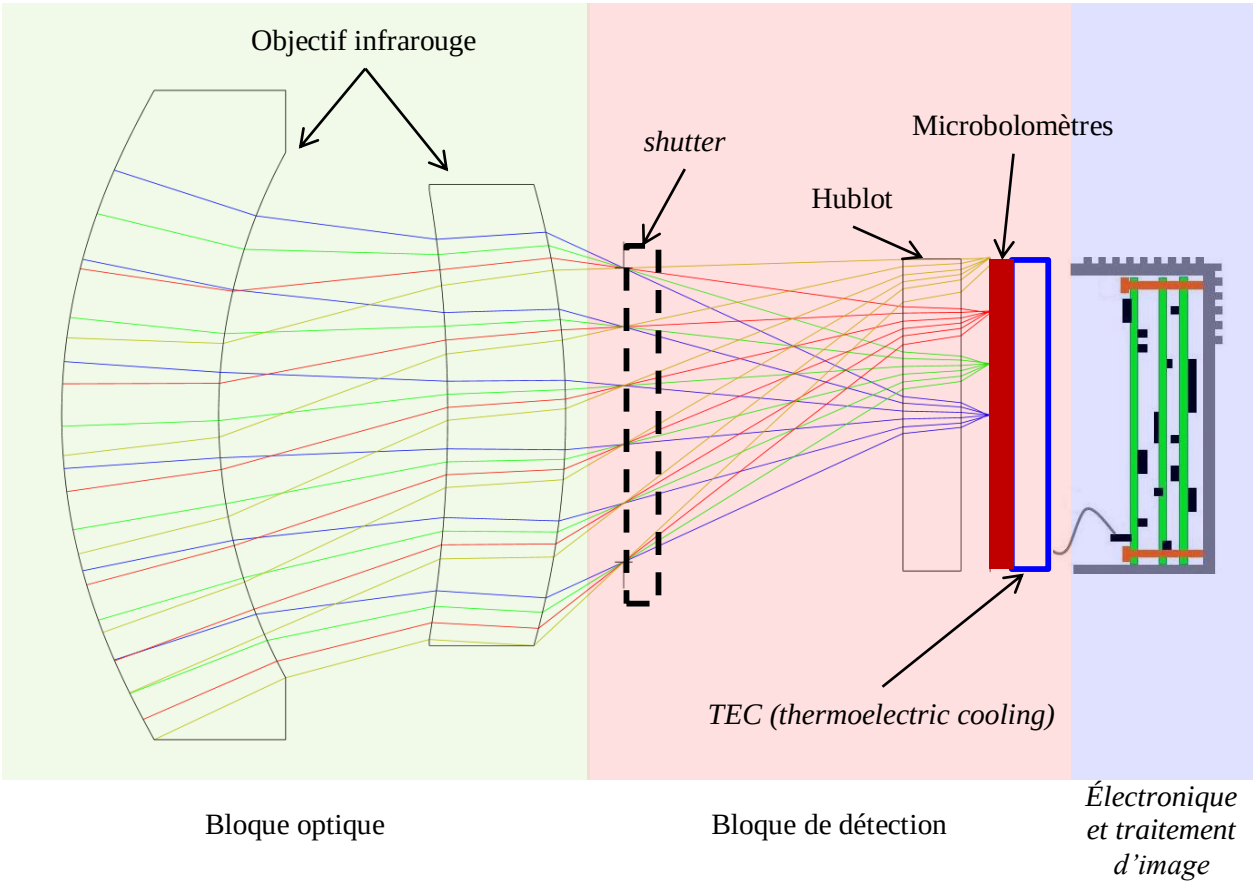


Figure 13. Illustration des éléments nécessaires à l'imagerie thermique par microbolomètres.

Le bloque optique est un autre élément qui coûte cher et des efforts sont également menés réduire son coût. Le matériau le plus performant pour réaliser des optiques est le germanium car son indice optique et sa transmission élevée sont avantageuses pour les opticiens. Mais ce matériau est cher et tend à être remplacé par des matériaux dérivés appelés chalcogénure. La particularité de l'imagerie thermique, est la nécessité d'avoir des optiques très « ouvertes » pour garantir une sensibilité thermique importante. L'ouverture *N* est le paramètre de l'optique à prendre en compte. Il correspond approxiamativement au rapport entre le diamètre de la pupille de sortie de l'optique et la distance focale. Les opticiens utilisent souvent l'expression : optique ouvertue à « f/*N* ». Une optique ouverte à f/1 signifie que la distance focale est égale au diamètre de la pupille de sortie de l'optique. Une optique ouverte à f/0.8 est plus ouverte qu'une optique ouverte à f/2, et par conséquent, une optique ouverte à f/0.8 permet d'obtenir une meilleure sensibilité thermique qu'une optique ouverte à f/2. Cependant la taille d'une optique ouverte à f/0.8 sera plus importante que celle d'une optique ouverte à f/2 pour une distance focale constante, l'optique ouverte à f/0.8 sera donc plus cher que l'optique ouverte à f/2 dans ce cas. Les fournisseurs de caméra thermique précisent donc souvent l'ouverture de l'optique lorsqu'ils donnent la sensibilité thermique de leur caméra.

Notons qu'il existe deux sous-filières microbolomètriques différenciables par le matériau transducteur utilisé. La société *Ulis* utilise du silicium amorphe (*a-Si*) alors que certaines matrices de bolomètres produites par la société américaine *FLIR* sont basées sur de l'oxyde de vanadium (*VOx*). Le *VOx,* grâce à son impédance élevée, possède un bruit de Johnson inférieur au *a-Si*. L'avantage du *a-Si* est sa compatibilité avec les processus de fabrication sur *wafer* en silicium. Ce qui laisse présager des possibilités de réduction de coût importante pour une production à large volume. De plus les progrès réalisés dans les techniques de dépôt des couches qui composent les pixels ont permis de diminuer les coûts de la filière *a-Si* en termes de résolution thermique.

Ulis propose désormais des matrices au format 80×80 au pas de 36 µm avec une résolution thermique de de 100 mK (@f/1 à 50 Hz) dans une gamme de température de -40°C à +85°C avec. *FLIR* propose sa gamme Lepton au format 80×60 pixels en *VOx* au pas de 17 µm avec une sensibilité thermique de 50 mK (détecteur seul) avec une correction de l'image *shutterless* dans une gamme de température de -10°C à +65°C. La fréquence d'acquisition annoncée est de 8.6 Hz. *Seek Thermal* propose une caméra thermique pour *smartphone* à 249 $ intégrant un détecteur *Raytheon* au format 206×156 pixels en *VOx* au pas de 12 µm.

1.3. Les marchés civils de l'imagerie thermique

1.3.1. L'automobile

Dans l'industrie automobile, l'imagerie thermique est essentiellement utilisée pour détecter des corps chauds tels que des piétons ou des animaux à l'extérieur du véhicule. L'équipementier *Autoliv* dès 2005 proposait un affichage de la scène routière filmée grâce à une caméra thermique microbolométrique [37]. Parfois décrié car le conducteur doit constamment regarder l'écran d'affichage de l'image thermique, des algorithmes de traitement d'images ont été ajoutés pour détecter des piétons et des animaux et ainsi les faire apparaitre plus distinctement sur l'image ou envisager un signal sonore en cas de danger imminent (cf. Figure 14). Il est également possible de permettre un freinage d'urgence au cas où le conducteur n'adapte pas sa trajectoire. Certains constructeurs premium proposent ce type de fonction comme *Mercedes*, *Rolls Royce*, *Cadillac*, *Audi* et *BMW*.

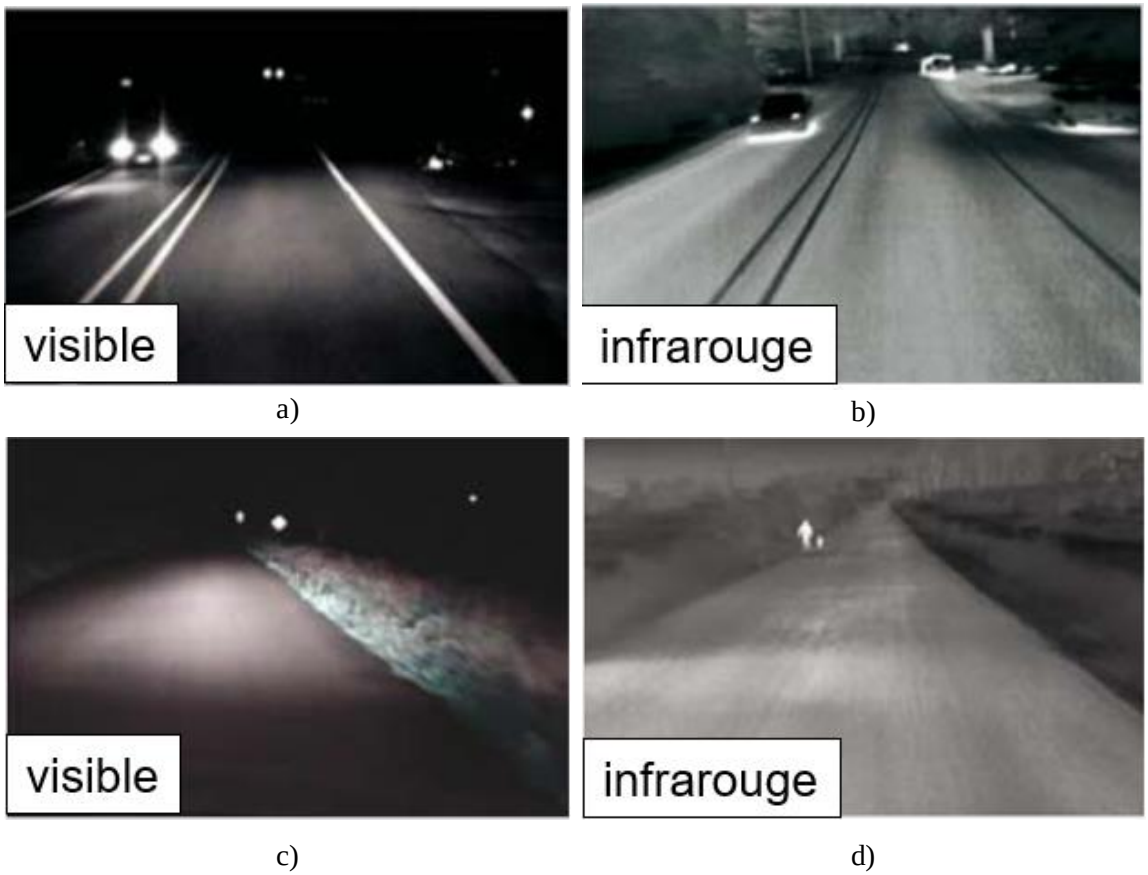


Figure 14. Illustration des systèmes de vision nocturne grâce à l'imagerie thermique. a) Image d'une scène routière dans le visible et b) image de la même scène routière le *LWIR*. c) Image visible d'une autre scène routière et d) image de la même en *LWIR*.

Un autre système couplant la détection de piétons et d'animaux à un projecteur lumineux dynamique permet de rendre visible pour le conducteur le danger potentiel sur le bord de la route (cf. Figure 15). La Figure 16 illustre l'intégration d'un tel système par *BMW*.

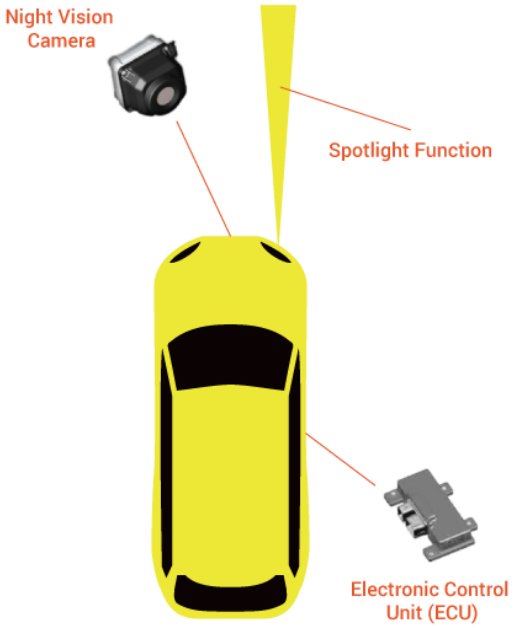


Figure 15. Système *Autoliv* de détection de piétons et d'animaux couplé à un système d'éclairement par projecteur lumineux visible [37].

Figure 16. Système d'éclairage dynamique de piéton proposé par *BMW* et *Autoliv* [38]. Une caméra *LWIR* détecte le point chaud, et un éclairage visible dynamique éclaire par impulsion le piéton.

Le projet *AWARE (all weather all roads enhanced vision)*, financé par le fond unique interministériel *FUI*, est une collaboration entre *Valeo*, *Sagem* et est piloté par l'entreprise française *Ulis* [39]. Le projet *AWARE* est focalisé sur la détection de piétons dans des conditions météorologiques difficiles. Quatre caméras sensibles à des bandes spectrales différentes sont testées : visibles (pixels *RGB* sensibles entre 0.45 et 0.65 µm), *NIR* (pixels sensibles entre 0.4 et 1 µm), *SWIR* (pixels sensibles entre 0.6 et 1.7 µm) et *LWIR*

(pixels sensibles entre 8 et 12 µm). Une base d'images est acquise dans des conditions météorologiques difficiles avec un (ou plusieurs) piéton(s) présent(s) dans le champ de vue des caméras. L'objectif est de déterminer la bande passante en longueur d'onde la plus appropriée aux conditions météorologiques. Afin de rendre le test le plus objectif possible, il a été décidé de demander à deux observateurs humains d'essayer de discerner visuellement un piéton dans les images. Par cette méthode on n'introduit pas de biais dus à des maturités différentes d'algorithme de détection de piétons (mais on introduit d'autres biais liés à la l'algorithme de *tone mapping*). Ce projet montre que les détecteurs thermiques microbolométriques non-refroidis sont les plus performants pour détecter des piétons en cas de brouillard dense. Un autre avantage est également mis en avant : les caméras thermiques ont des résultats plus reproductibles que les autres caméras pour les tâches de détection et de reconnaissance, lorsque les conditions météorologiques sont très variées. L'inconvénient de l'imagerie thermique est son incapacité à reconnaitre des panneaux de circulation ainsi que les marquages au sol.

Nous venons de citer quelques exemples d'utilisation de caméra thermique à l'extéreur du véhicule. D'autres applications, à l'intérieur du véhicule, peuvent également tirer parti des avantages des caméras thermqiues.

L'interface homme machine est un domaine actif dans l'industrie automobile. Les constructeurs font des efforts pour rendre intuitives les fonctionnalités annexes du véhicule telles que le téléphone main libre, la radio ou le navigateur. La technologie *touch screen* par dalle tactile capacitive (ou résistive) identique à celle présente sur un grand nombre de *smartphones* et de tablettes s'est largement imposée. Ce type d'interface implique que le regard du conducteur quitte la route. Pour contrer cela, la commande par geste, également connue sous le terme anglophone *gesture controls*, peut être une solution. Une caméra temps de vol (aussi connue sous le nom de caméra *TOF* pour *time of flight*) positionnée entre le rétroviseur central et le plafonnier a été intégrée par *BMW* dans ses habitacles pour les fonctionnalités *gesture controls* [40]. Des alternatives techniques de commande par gestes utilisant une caméra thermique ont été évoquées dans la littérature scientifique dans un contexte qui n'est pas spécifique à l'automobile. Une caméra thermique qui filme une surface (qui peut être différente de la tablette d'interface) peut détecter aisément un ou plusieurs doigts qui touchent ladite surface grâce à la trace résiduelle de chaleur (cf. Figure 17) [41,42]. De nouvelles possibilités de reconnaissance de geste sont donc potentiellement accessibles avec des caméras thermiques.

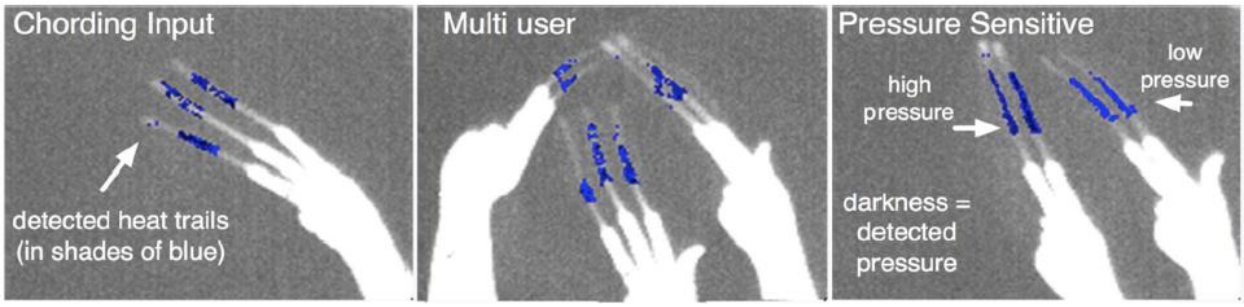


Figure 17. Trace résiduelle de chaleur par toucher sur une surface (qui peut être différente de celle de la tablette d'interface) [41].

Le laboratoire de recherche en électronique du groupe Volkswagen, situé en Californie, est partenaire du *LISA* (*Laboratory for intelligent and safe automobiles*). M. Trivedi, directeur du *LISA*, est à l'origine de nombreux travaux sur la surveillance du conducteur. Une *VW Passat* équipée de nombreuses caméras à l'intérieur de l'habitacle constitue un banc de test appelé *LISA-P*. Des caméras thermiques microbolométriques sont disponibles parmi le panel de caméras. L'une d'elle a été utilisée pour développer un système d'airbag intelligent pour le passager avant [32]. En fonction de la posture de celui-ci, l'airbag peut être neutralisé pour éviter des blessures importantes (cf. Figure 18). Le système d'estimation de la pose du passager ainsi créé est comparé à un système stéréovision *NIR* actif (utilisation de Leds dans le *NIR* pour que le système fonctionne la nuit). Les auteurs obtiennent de meilleurs résultats avec le système stéréovision qu'avec la caméra thermique notamment dans les situations où le participant porte un chapeau ou lorsqu'il tourne la tête de telle sorte que seul les cheveux sont visibles par la caméra thermique. Notons que ces travaux ne comparent pas uniquement les capacités intrinsèques d'un système stéréovision par rapport à une caméra thermique. En effet, le choix de l'algorithme de traitement d'image entre également en compte.

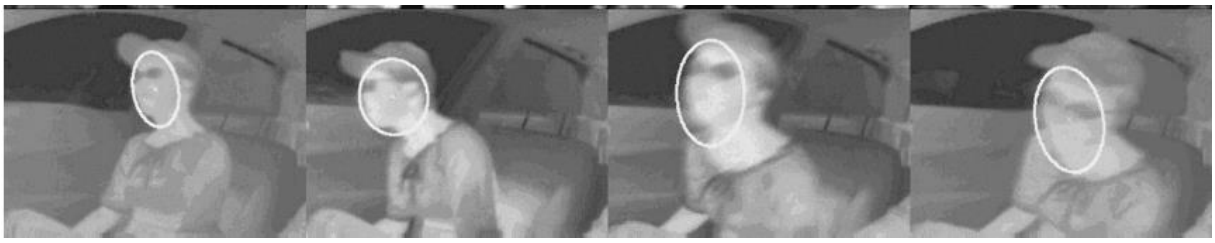


Figure 18. Système d'airbag intelligent [32]. Une caméra *LWIR* est utilisée pour détecter le visage. La taille et la position de l'ellipse sont utilisées pour déterminer si le passager est dans une zone dangereuse en cas de déclenchement d'airbag.

Enfin l'Université Technique de Munich a exprimé le besoin d'adapter la localisation des alertes dans de détection de piétons [30]. Les auteurs préconisent d'afficher une alerte sur le pare-brise grâce à un *HUD*. Cette alerte doit être localisée dans la direction définie par une droite imaginaire qui relie le piéton et la tête du conducteur. Les capteurs à l'extérieur du véhicule permettent de localiser le piéton, et leurs travaux proposent de localiser, grâce à une caméra thermique, la position de la tête du conducteur. Une estimation de la position 3D du visage est donc nécessaire. Les auteurs réduisent le problème à l'estimation de la position 2D du visage dans le plan du damier de la Figure 19 car ils estimment que la totalité des conducteurs adoptent une position identique selon l'axe perpendiculaire au plan du damier. Ainsi, une caméra, dont l'axe optique est positionnée perpendiculairement au plan du damier, permet d'estimer la position 2D du visage dans le plan du damier. La raison qui pousse les auteurs à utiliser une caméra thermique est la robustesse aux variations d'illumination, et l'aspect de sécurité oculaire grâce à la passivité du système basé sur une caméra thermique.

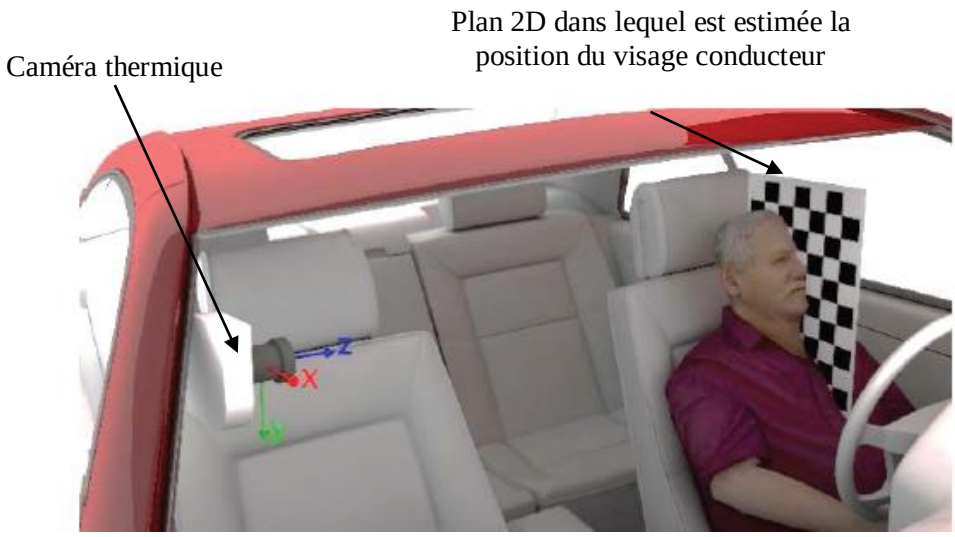


Figure 19. Estimation de la position 2D du visage du conducteur par caméra thermique [30].

Le coût d'une caméra thermique n'est pas encore justifié pour l'industrie automobile car il n'a pas été démontré que des informations critiques pouvaient être obtenues de manière extrêmement fiable. Ainsi leur prix est souvent comparé à celui des caméras visibles, beaucoup plus faible. Des projets en partenariat avec des acteurs du domaine de l'imagerie thermique non-refroidie, et des acteurs de l'industrie domotique, visant à réduire le coût des caméras thermiques microbolométriques, ont été menés comme nous allons le voir dans la section suivante.

1.3.2. La domotique

Le domaine de la domotique utilise depuis bien longtemps le rayonnement thermique via les détecteurs pyroélectriques pour détecter un véhicule devant une porte de garage ou un humain dans une pièce pour gérer automatiquement l'allumage des luminaires. Ce type de capteurs est bien souvent mono pixel et ne détecte que la présence d'objets en mouvement.

De nouveaux besoins ont été formulés suite à l'émergence du bâtiment intelligent, connecté, sécuritaire et économe en énergie. L'objet d'intérêt du point de vue d'un système de surveillance domotique est toujours l'être humain. Mais désormais, il est nécessaire de compter plusieurs personnes dans une même pièce et même d'estimer leurs activités. Ce qui demande des capteurs avec plusieurs pixels pour résoudre les objets d'intérêt. Ces nouveaux besoins sont donc relativement proches des besoins de surveillance du conducteur (ou plus généralement de l'habitacle du véhicule).

Le projet *MIRTIC (mirco retina thermal infrared circuit)*, terminé récemment, a été porté par plusieurs entreprises, notamment par la société *Schneider Electric*, acteur majeur dans l'industrie domotique et par le *CEA-Leti*, acteur majeur de la recherche française dans le domaine de la micro-électronique et par l'entreprise *Ulis* qui a piloté ce projet. L'objectif du projet *MIRTIC* est de concevoir un imageur à bas cout et nécessitant une faible consommation énergétique par rapport aux imageurs historiquement développés par la société Ulis et réservés au domaine militaire. Ce type d'imageur serait utilisé pour optimiser le chauffage (ou la climatisation) d'un bâtiment ainsi que son éclairage, en fonction de l'activité humaine.

D'autres applications telles que le contrôle d'accès à certaines zones ou la sécurité autour d'un véhicule sont également évoquées. L'effort mené dans ce projet consiste à réduire le coût des imageurs au format 80×80 pixels développés par la société *Ulis*. La particularité des imageurs microbolométriques est la mise sous vide des pixels. Cela permet d'atteindre des performances élevées en résolution thermique par rapport aux autres filières de l'infrarouge thermique non-refroidi (pyroélectrique et thermopile). Jusqu'alors, le vide était réalisé sur l'ensemble de la matrice. A l'occasion de ce projet, la société *Ulis* a proposé d'encapsuler séparément sous vide les pixels ce qui conduit à la réduction des coûts à iso performance de l'imageur. Cette technologie, nommée *PLP* pour *pixel level packaging* dans la littérature anglophone a été développée en étroite collaboration avec le *CEA-Leti* [43]. L'imageur conçut lors de ce projet est nommée Micro80P.

Un projet de surveillance des personnes âgées a été réalisé sur la base de la technologie thermopile. Il s'agit du projet *FUI E-monitorage* porté par la société *Legrand* qui inclut partiellement les travaux de thèse CIFRE de T. Guettari (2014) [44]. Cette thèse propose d'associer un algorithme *K-means* à un pixel thermopile pour détecter la présence d'une personne dans un lit, dans le cadre des établissements d'hébergement pour personnes âgées dépendantes (*EHPAD*). Au-delà de la détection de présence, ces travaux montrent qu'il est possible d'estimer le niveau du sommeil à partir de la mesure de température du pixel thermopile. Ce lien entre sommeil et température est détaillé dans la section suivante.

Que cela soit dans le domaine de la domotique ou de l'automobile, un certain nombre de projets et d'applications commerciales impliquant des caméras thermiques ont été menés. A chaque fois, le corps vivant et plus particulièrement celui de l'être humain est l'objet à détecter. Notre projet, la surveillance du conducteur, semble être adressable par ce type de caméra, au moins parce qu'il semble relativement aisé de discerner un être humain de manière robuste. La section suivante apporte des perspectives plus audacieuses mais tout à fait sérieuses pour la surveillance du conducteur par caméra thermique.

1.4. Des perspectives : surveillance des paramètres physiologiques

1.4.1. Somnolence et imagerie thermique

Le système nerveux sympathique contrôle un grand nombre d'actions inconscientes. Il est notamment responsable des alternances entre le sommeil et l'éveil. A l'approche du sommeil, la sécrétion de mélatonine augmente ainsi que la pression sanguine. Les vaisseaux se dilatent et la chaleur interne du corps est dissipée par la peau. Le résultat de ce processus est une baisse de la température interne du corps. Les mécanismes de causes à effet sont encore mal connus [45]. Cependant il est admis par la communauté scientifique de la chronobiologie qu'une augmentation de la température de certaines zones externes richement vascularisées est observable à l'approche d'une phase de sommeil attendue (c'est-à-dire lorsque le sujet ne lutte pas contre le sommeil). Une étude évalue l'augmentation de la température des doigts avant une période de sommeil [46]. Ce test a été conduit sur 14 participants d'âge et de sexe différents avant des périodes de sommeil. La dérivée temporelle de la température atteint un maximum de 0.8°C/min en moyenne (avec un écart type de 0.09°C) avant l'endormissement. La plage totale d'augmentation de la température varie entre 1 et 3°C.

Un indicateur de déclin du niveau d'éveil à partir de la mesure sans contact de la température de la zone nasale a été proposé [47]. Une caméra thermique non-refroidie est utilisée pour la thermographie nasale. Concernant la vérité terrain sur l'état d'éveil, un électroencéphalogramme (EEG, pour tester l'atténuation des ondes alphas), une échelle subjective (on demande au sujet d'indiquer sur une échelle son niveau d'éveil) sont utilisés. Le test porte sur quatre sujets de 22 ans à qui on fait écouter pendant six minutes une musique calme ou dynamique afin de provoquer des états d'éveils différents. Leurs travaux montrent qu'une augmentation de la température de la zone péri-nasale survient lorsque le niveau d'éveil diminue. Cependant l'amplitude de ces variations d'un participant à l'autre n'est pas négligeable. De plus les auteurs soulignent le fait que leurs tests ont été conduits dans un laboratoire contrôlé en température. Dans un habitacle automobile, une illumination directe des rayons solaires sur la peau pourrait perturber les mesures. Pour contrer cela, nous pourrions envisager de normaliser la température du nez par la température moyenne du visage entier.

Remarque : il a été question en début de thèse d'une collaboration avec K. Kraüchi de la clinique psychiatrique universitaire de Bâle. Ce chercheur en chronobiologie a réalisé de nombreux travaux sur le lien entre le sommeil et la température de la peau [48]. Pour des raisons de complexité administrative (K. Kraüchi a pris sa retraite au début de cette thèse) cette collaboration n'a malheureusement pas pu aboutir.

D'autres paramètres physiologiques sont communément utilisés pour estimer le niveau de fatigue. Le rythme cardiaque et le rythme respiratoire en font partie. Les enjeux de l'industrie automobile sont de mesurer ces paramètres précisément sans générer de gêne pour le conducteur afin de rendre le système acceptable. Des travaux de 2005 proposent d'estimer les battements cardiaques d'un individu grâce à une caméra thermique [49]. Une caméra de haute qualité a été utilisée : une caméra *MWIR FLIR Indigo Phoenix* (détecteur quantique *InSb*). Les vaisseaux sanguins les plus proches de la surface de la peau chauffent celle-ci au rythme des battements cardiaques. La Figure 20 illustre les zones utilisées pour estimer le rythme cardiaque. Il s'agit de la veine située sur la partie supérieure de l'avant bras, la carotide externe ou le complexe de veines et d'artères situé sur la tempe. Le rythme cardiaque est estimé avec une grande précision : le coefficient de corrélation (coefficient de Pearson) avec la vérité terrain vaut 0.994. Bien que certains travaux montrent que le rythme cardiaque (*HR*) peut être un bon candidat pour créer un indicateur de fatigue, la variabilité du rythme cardiaque est plus souvent utilisée [50]. Dans la limite de nos connaissances, la question de la possibilité de mesurer le *HRV* avec une caméra thermique non-refroidie reste en suspens.

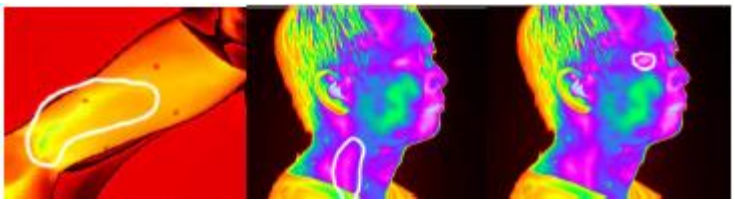
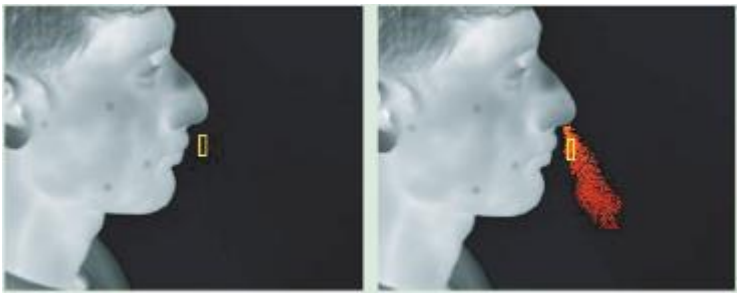


Figure 20. Zones privilégiées pour l'estimation du battement cardiaque [49].

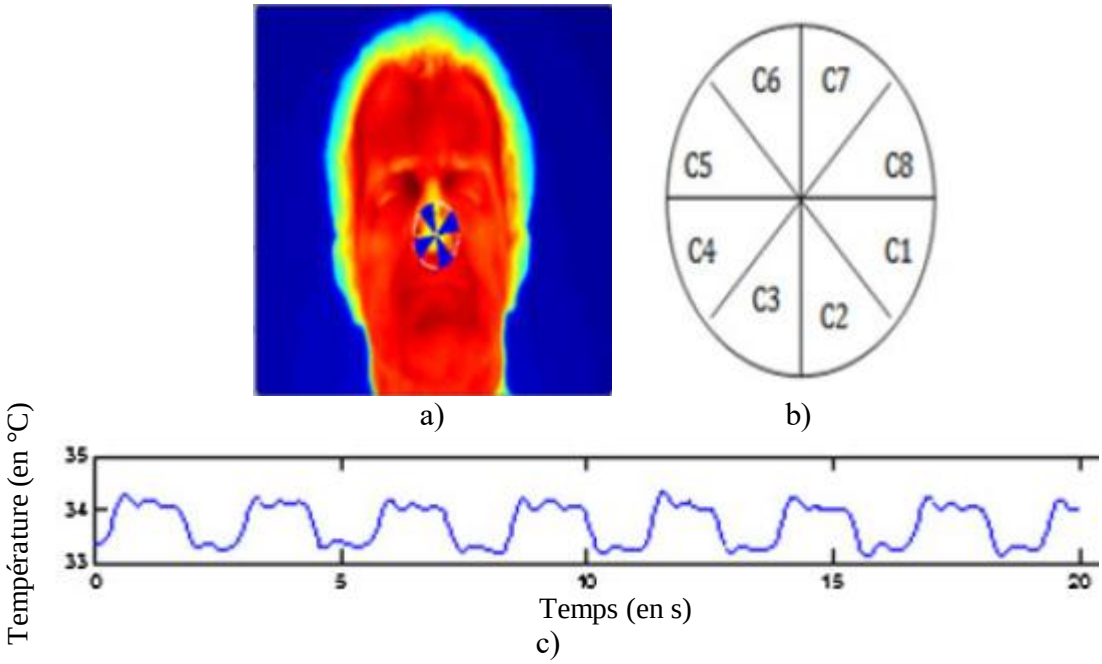
Au sein du même groupe de recherche, des travaux menés par R. Murthy et I. Pavlidis permettent d'estimer la fréquence respiratoire [51]. La Là aussi, la même caméra de haute qualité a été utilisée. La Figure 21 illustre la zone sous le nez qui est étudiée pour l'évaluation du rythme respiratoire.



a) b)

Figure 21. illustration de la zone d'intérêt permettant d'évaluer le rythme respiratoire [51]. a) L'individu n'expire pas. b) L'individu expire de l'aire.

Les auteurs de la référence [52] proposent également d'estimer le rythme respiratoire en analysant une zone proche du nez. Mais une caméra non-refroidie (*LWIR*) est utilisée à la place d'une caméra *MWIR*. La Figure 22 illustre les neufs zones analysées et la moyenne spatiale des pixels appartenant à la zone C7. Ces premiers tests montre que la variation à suivre est de l'ordre du degré Celsius.



a) b) c)

Figure 22. Illustration des zones analysées pour l'évaluation du rythme respiratoire [52]. a) Image thermique et zones analysées. b) Numérotation des zones analysées. Les pixels appartenant à une zone sont moyennés spatialement. Les moyennes des zones sont suivies temporellement. c) Moyenne des pixels de la zone C7 en fonction du temps.

La température, le rythme respiratoire (voire le rythme cardiaque) sont des paramètres physiologiques mesurables par une caméra thermique pour l'évaluation du niveau de fatigue. Certains travaux propose même d'estimer certains de ces paramètres à l'aide d'une caméra thermique non-refroidie. Le point commun de ces travaux est qu'ils nécessitent d'être capable d'effectuer les tâches suivantes :

- un algorithme de suivi (*tracking*) du visage en imagerie thermique pour faciliter le suivi des zones d'intérêt,
- un étalonnage permettant de rendre la mesure de température par caméra thermique possible.

Dans la section suivante, d'autres paramètres physiologiques pouvant être surveillés par un imageur thermique seront abordés. Les besoins techniques (algorithme de *tracking* et étalonnage thermique) sont approximativement identiques à ceux de l'évaluation du niveau de fatigue.

1.4.2. Etat mental et imagerie thermique

Les émotions humaines auraient tendances à faire varier le flux sanguin au niveau du visage, notamment au niveau du front. La conséquence de cela est une dissipation de la chaleur à travers la peau. Un état de stress ou une charge mentale importante pourrait ainsi être détecté grâce à un imageur thermique. Voici quelques exemples de travaux scientifiques qui appuient cette hypothèse.

La relation entre la charge mentale d'un opérateur (tel qu'un conducteur) et la température du visage est explorée [53]. Douze participants réalisent des tâches qui nécessitent une concentration qui peut-être basse, moyenne ou importante. Une caméra thermique filme le visage et sept zones sont différenciées (front, nez, menton, joue droite, joue gauche, yeux) et suivies (*trackées*) automatiquement grâce à un *tracker 6-DOF (*de l'anglais *degrees of freedom) InterSense-900*. Les températures moyennes sont utilisées comme les données d'entrée d'un réseau de neurones dont l'objectif est d'estimer la charge mentale des participants. La méthode permet d'estimer correctement la charge mentale dans 81% des cas lorsque le réseau est entrainé sur 12 personnes. Le taux de bonne détection peut atteindre 98,9% si le réseau de neurone est entrainé spécifiquement pour une personne. Les auteurs soulignent le fait que la signature thermique de la charge mentale est variable d'un individu à l'autre.

Il y a aussi des propositions de création d'un indicateur de stress basé sur la mesure de la température du front [54]. Cet indicateur est comparé à une méthode d'évaluation du stress bien établie dans le domaine médical : la dépense d'énergie mesurée par calorimétrie indirecte (typiquement le participant porte un masque relié à une machine qui analyse la quantité de dioxygène expirée et la quantité d'oxygène consommée). Les tests montrent une bonne corrélation entre l'indicateur basé sur l'image thermique et la vérité terrain fournie par calorimétrie indirecte. Pour détecter les variations thermiques, une caméra *MWIR FLIR Indigo Phoenix* (détecteur quantique *InSb*) est utilisée.

La température de la zone sus-orbitale du visage (c'est-à-dire le front) est également exploitée [55] pour estimer la charge mentale liée à l'utilisation d'un *smartphone* (message écrit ou conversation téléphonique) d'un conducteur. Grâce à une expérience de simulation de conduite (cf. Figure 23), il est démontré que le fait d'effectuer une tâche supplémentaire telle qu'écrire un message sur un *smartphone* ou avoir une conversation téléphonique augmente la dérivée temporelle de la température de la zone sus-orbitaire (la dérivée de la température est calculée sur une fenêtre de 20 s). Pour détecter ces variations, une caméra *MWIR FLIR SC6000* (détecteur quantique *InSb*) de sensibilité thermique 25mK est utilisée. Les auteurs constatent également une dégradation des performances de conduite (non maintien de la vitesse et de la position latérale du véhicule) lors des phases de surcharge cognitive.

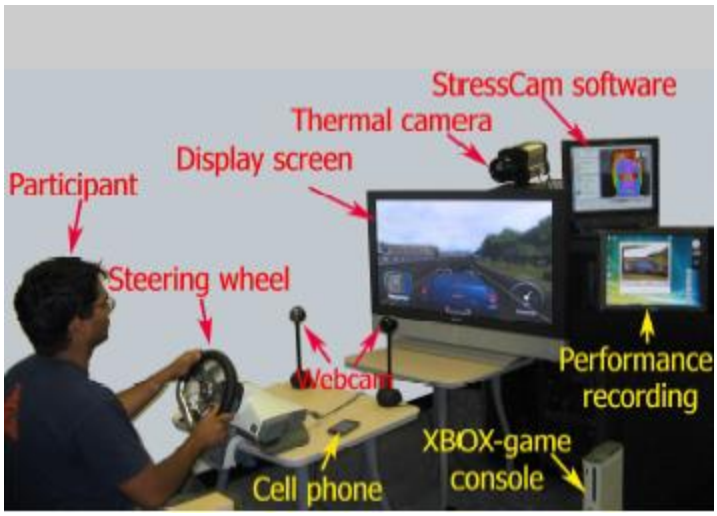


Figure 23. Expérience menée dans la référence [55] pour estimer la (sur)charge mentale dans une situation de simulation de conduite.

Précisons que la zone sus-orbitale est initialisée manuellement puis suivie (c'est-à-dire *trackée*) automatiquement grâce à un algorithme proposé en 2013 par Y. Zhou *& al* [56]. La performance de *tracking* de la *ROI* (de l'anglais *region of interest*) impacte directement l'évaluation de l'état mental du participant. Y. Zhou *& al* [56] soulignent le fait que l'apparence thermique du visage peut être impactée par des critères physiologiques comme une allergie. Il n'est pas donc pas aisé de *tracker* une petite zone du visage en se basant sur un modèle d'apparence locale. Les auteurs proposent alors un modèle probabiliste qui initialise un filtre particulaire. Le filtre particulaire réalise un lissage temporel et spatial. Leur méthode est robuste aux larges variations de pose et aux variations physiologiques du visage (cf. Figure 25).

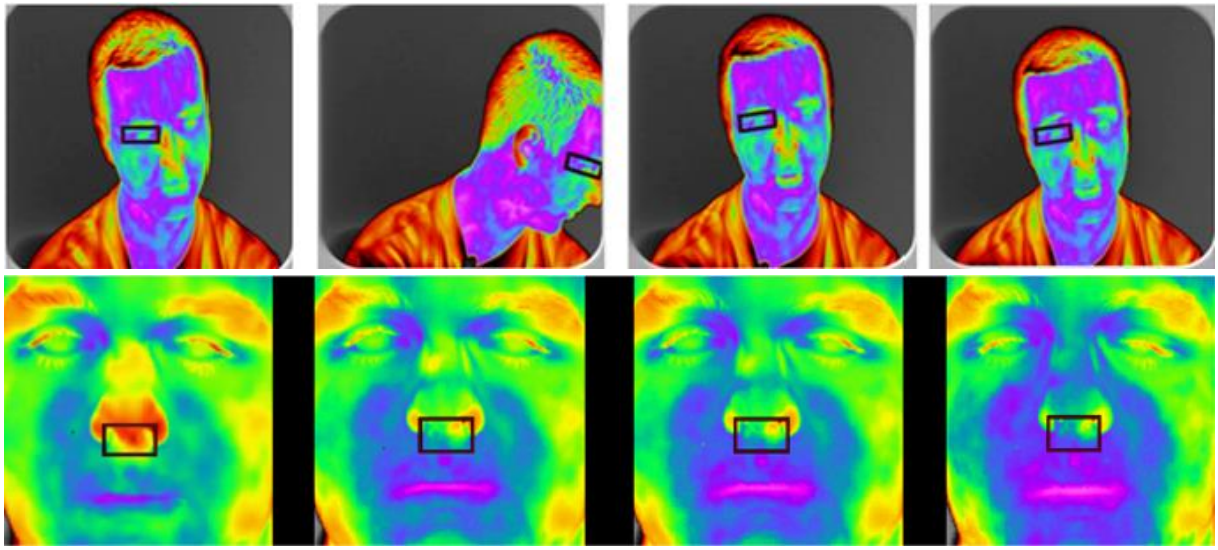


Figure 25. Illustration de la méthode de *tracking* d'une *ROI* détaillée dans la référence [56]

Enfin, un article de I. Pavlidis *& al* paru dans *Nature Scientific Reports*, détaille les effets de différentes sources de stress (émotionnel, cognitif ou moteur) sur la performance de conduite en simulation sur un panel de 59 participants [57]. Pour établir l'état de stress les auteurs utilisent une caméra thermique basée sur une matice de microbolométres, il s'agit de la caméra *FLIR TAU 640*. La zone de température révélatrice d'un état de stress est située entre le nez et la bouche (Figure 24). La difficulté est d'analyser l'augmentation de température de cette zone en sachant que la respiration provoque une variation périodique de température.

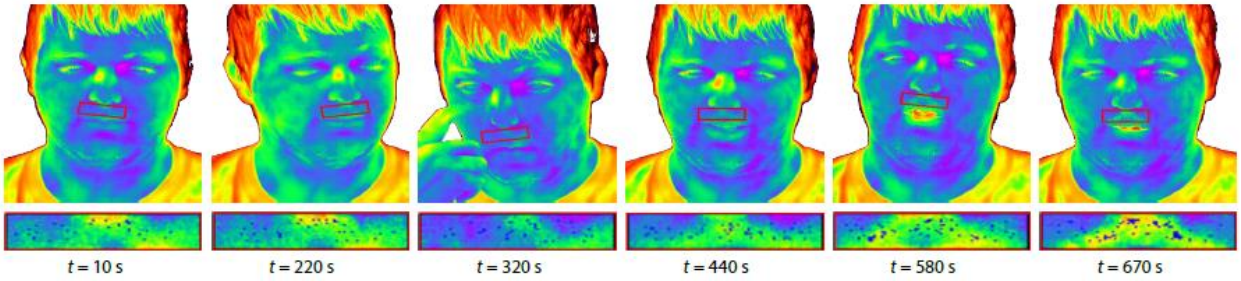


Figure 24. Illustration de l'augmentation de la température de la zone péri-nasale lors d'une situation de stress [57].

1.5. Les algorithmes d'estimation de la pose du visage

La pose du visage est la combinaison de la position et de l'orientation du repère attaché au visage par rapport à un repère monde qui est dans notre cas, un repère attaché au véhicule. Les trois angles d'Euler qui définissent l'orientation du repère objet sont représenté sur la Figure 26. Nous utiliserons dans ce manuscrit leur dénomination anglophone. Les angles *yaw* (lacet), *pitch* (tangage), et *roll* (roulis) désignent les angles autour des axes respectifs $\boldsymbol{Y}^O$, $\boldsymbol{X}^O$ et $\boldsymbol{Z}^O$.

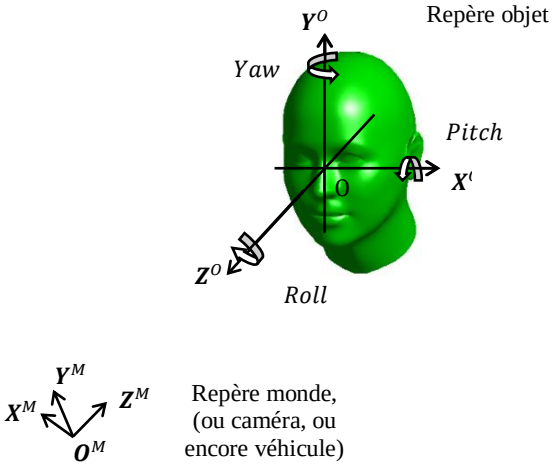


Figure 26. Illustration des repères monde et objet, ainsi que des angles *yaw*, *pitch* et *roll*.

L'estimation de la pose du visage est point essentiel car elle peut être exploitée directement pour tirer des informations relatives à la direction du regard ou aux hochements de la tête. Il est également indispensable dans les applications où l'on souhaite extraire des paramètres physiologiques du visage. En effet, à notre connaissance, bien que des algorithmes de suivi (*tracking*) de tissues du visage ont été développés, l'initialisation de la zone à *tracker* se fait toujours manuellement [56]. Lors d'un mouvement ample du visage, si la zone à *tracker* devient invisible, le système ne fonctionne plus. La combinaison d'un algorithme d'estimation de la pose avec un algorithme de suivi de tissues, permettrait d'être robuste aux mouvements amples du visage.

Des algorithmes ont été développés dans le visible/*NIR*, et nous les abordons à la section 1.5.1. La nature de l'image *LWIR* étant différente de celle de l'image visible/*NIR*, rien ne garantit que ceux-ci sont directement adaptables à l'imagerie thermique. Des travaux spécifiquent proposent d'estimer la pose, ou une partie de la pose, à partir d'images thermique. Nous les abordons à la section 1.5.2.

1.5.1. Les algorithmes d'estimation de la pose dans le visible/*NIR*

M. Chutorian et M. Trivedi proposent une revue des algorithmes d'estimation de la pose [58]. Un classement très complet, en fonction des caractéristiques des différentes méthodes, y est réalisé. Nous adopterons dans cette sous-section un classement beaucoup plus simple que celui de M. Chutorian et M.

Trivedi. En effet, nous distinguerons seulement deux types d'algorithmes : ceux basés sur une méthode « globale » et ceux basés sur méthode « locale ».

- Méthode « globale » : tous les pixels du visage, c'est-à-dire les pixels de l'image qui représente le visage du conducteur, sont utilisés pour estimer la pose du visage.
- Méthode « locale » : seuls les points d'intérêts sont utilisés pour estimer la pose du visage. Un point d'intérêt est une petite zone de l'image, par exemple une fenêtre de 16×16 pixels, qui présente des caractéristiques remarquables. Une caractérisitique remarquable est par exemple un gradient de l'image important le long d'un ou plusieurs axes.

Citons une première méthode « globale », qui s'appuie sur une base d'images de visages différemment orientés et labellisée [59]. A partir d'une seule image d'un visage représentée au centre de la Figure 27, les auteurs proposent de générer les huit autres images virtuelles du visage orienté différement. Pour cela, ils proposent d'appliquer un modèle de déformation. Ensuite, un score reposant sur une corrélation normalisée est calculé pour estimer la pose dans l'image réelle (c'est-à-dire l'image acquise par la caméra). On obtient ainsi une estimation grossière de la pose du visage dans le cadre d'un algorithme de reconnaissance faciale.

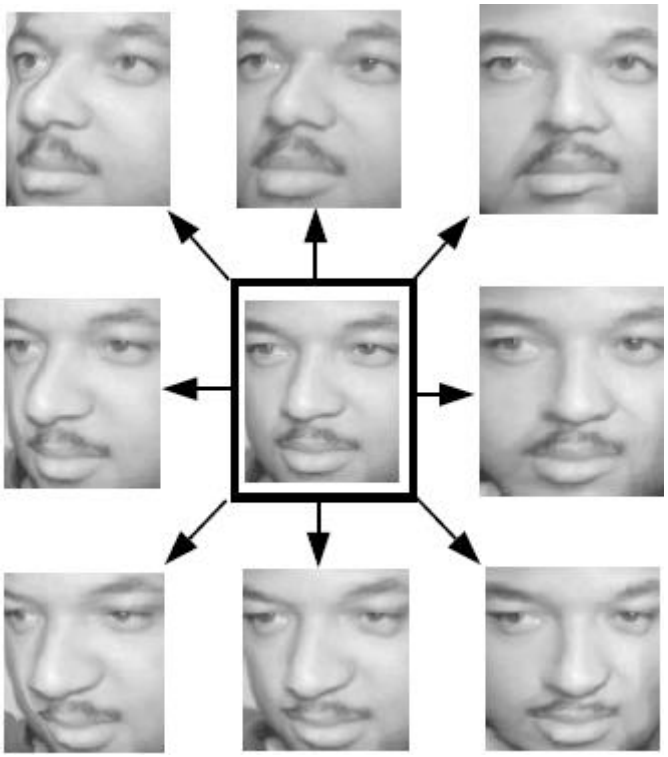
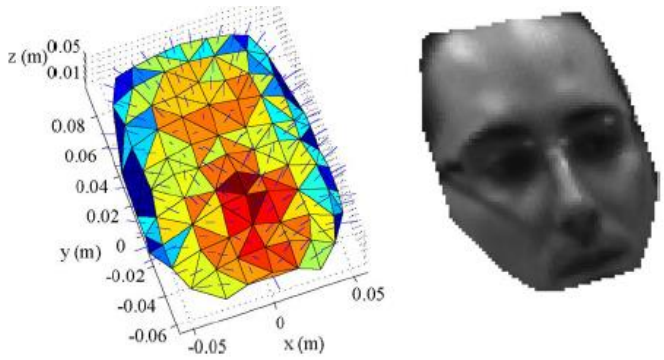


Figure 27. Illustration de la création d'une base d'images virtuelles. Au centre, il s'agit de l'image réelle. Les huit autres images ont été synthétisées et représentent des orientations spécifiques du visage [59].

Une autre méthode « globale » a été proposée par M. Chutorian et M. Trivedi [60]. Ils proposent de détection du visage et d'estimation de la pose complètement automatique. Le visage est détecté grâce à un algorithme basé sur un apprentissage supervisé et une décomposition en ondelette de Haar. Une estimation



a) b)

Figure 28. Maillage 3D texturé utilisé pour le suivi de la pose [60]. a) maillage 3D. b) maillage 3D texturé.

initiale de la pose est opérée grâce à trois *SVRs* (*support vector regressions*), chacun entrainé pour reconnaitre une orientation (c'est-à-dire *yaw*, *pitch* et *roll*). L'histogramme local du gradient orienté (*LGO* pour *localized gradient orientation*) est utilisé en entrée des *SVRs*. La pose initiale est utilisée pour aligner le maillage 3D, représenté sur la Figure 28 a), avec une image réelle et ainsi permettre le plaquage de la texture. Le maillage 3D texturé est représenté sur Figure 28 b). Ensuite, grâce à un environnement virtuel, différentes projections du modèle 3D sont comparées à l'image réelle grâce à une corrélation normalisée. Un filtrage particulaire tire parti de l'accumulation temporelle des mesures de la pose pour prédire la pose de l'image suivante. Ainsi, il est possible de réduire le nombre de poses virtuelles à comparer à l'image réelle. L'algorithme final fonctionne à 30 Hz.

L. P. Morency *& al* proposent également une méthode « globale » complétement automatique et illustrée sur la Figure 29 [61]. La pose est estimée de trois manières différentes représentées par les trois flèches oranges de la Figure 29 :

1. Une estimation grossière de la pose est effectuée grâce à des algorithmes de détection du visage de type *Viola-Jones*. La méthode *Viola-Jones* est algorithme de détection d'objet basé sur un apprentissage supervisé. Cette méthode est « globale » car tous les pixels du visages sont ulitisés.
2. La pose différentielle est estimée entre l'image actuelle et l'image précédente.
3. La pose différentielle est estimée entre l'image actuelle et une sélection d'images clefs choisies grâce à un filtre de Kalman. Un filtre de Kalman utilise l'accumulation temporelle des estimations pour prédire l'estimation de la pose suivante.

L'étape 1 fonctionne pour un grand nombre d'individu, mais permet seulement une estimation grossière de la pose. L'étape 3 est plus fine mais nécessite une base d'images clefs, spécifique à l'utilisateur. Celle-ci se crée et s'enrichit en fonctionnement « en ligne » grâce aux estimations de la pose, comme le montre les flèches vertes sur la Figure 29. Pour les étapes 2 et 3, un algorithme appelé *normal flow constraint (NFC*), permet d'estimer l'évolution de la pose.

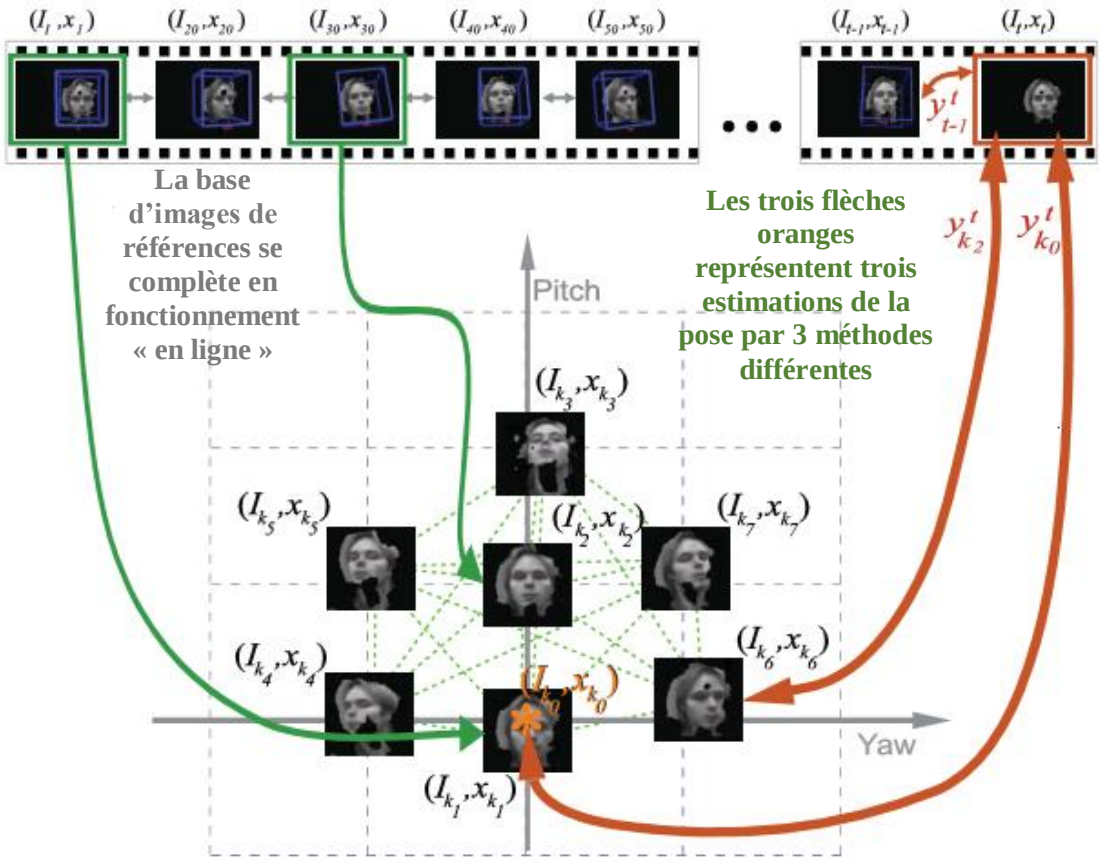
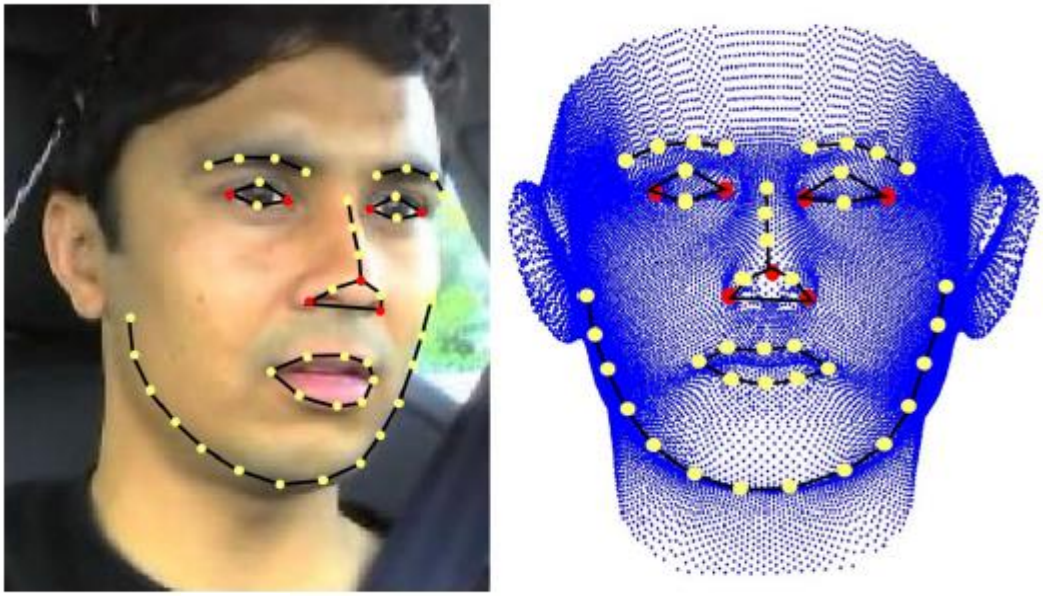


Figure 29. Illustration de l'algorithme « globale » développé par L. P. Morency *& al* [61]. Les flèches vertes symbolisent le fait que la base d'images de synthèses se complète au fur et à mesure que des poses du visage apparaissent. Les flèches oranges illustrent les trois manières d'estimer la pose.

Abordons désormais quelques exemples de méthodes locales. Dans la référence [62], les auteurs suggèrent de détecter des points caractéristiques. Les points d'intérêt auxquels on donne une sémantique sont des points caractéristiques. les coins intérieurs et extérieurs des yeux ainsi que la pointe du nez sont les points caractéristiques utilisés dans la référence [62]. L'angle entre la ligne des yeux et l'horizontale permet d'estimer l'angle *roll*. L'angle *yaw* est estimé à partir de la différence de taille entre les deux yeux. Enfin l'angle *pitch* est estimé en analysant la taille du segment normal à la ligne des yeux qui rejoint la pointe du nez. Cet algorithme sous-entend que des points spécifiques du visage peuvent être détectés et reconnus automatiquement.

Les méthodes algorithmiques récentes réalisent cette reconnaissance des points caractéristiques automatiquement [63]. En considérant que tous les visages ont approximativement tous la même forme, la position 3D de ces points caractéristiques est facilement mis en correspondance avec leur position 2D sur l'image comme cela est illustré sur la Figure 30. Sur la Figure 30 a), on peut voir les points caractéristiques 2D sur l'image réelle, et sur la Figure 30 b), on peut voir leurs corresponts 3D. Il est mentionné deux inconvénients à ce type de méthode. Le premier est qu'en cas de large rotation du visage, certains points caractéristiques ne sont plus visibles,-l'estimation de la pose devient impossible. Le second est que la reconnaissance des points d'intérêts nécessite un apprentissage supervisé. La variation inter-

individus de l'apparence des yeux, de la bouche et du nez (etc…) devient une difficulté à gérer et à intégrer dans la base de données.



a) b)

Figure 30. Un exemple de méthode d'estimation de la pose du visage implémentée en 2014 [63]. a) Les points caractéristiques en 2D sont détectés sur l'image réelle. b) La position 3D correspondante est évaluée grâce à un modèle générique. Seulement 4 correspondances 2D-3D suffisent pour estimer la pose relative du visage par rapport à la caméra.

D'autres méthodes locales s'appuient sur des points non-spécifiques. Les points non-spécifiques sont des points d'intérêt auxquels on ne donne pas de sémantique. V. Lepetit et P. Fua les appellent *natural points* dans la référence [63-64]. H. Wang *& al* utilisent ces points dans la référence [65]. Ils tirent bénéfice d'un modèle 3D texturé du visage pour créer des images de synthèse du visage. Puis, ils mettent en correspondance des points d'intérêt entre les images de synthèse et l'image réelle. Sur la première ligne de la Figure 31, on peut voir trois images de synthèses. Sur la seconde ligne, on peut voir des images réelles. Des correspondences de points d'intérêts non-spécifiques sont établies entre l'image de synthèse d'une colonne et l'image réelle de la même colonne. Nous employons une technique similaire qui sera détaillée au chapitre 5 car nous pensons qu'il n'est pas aussi aisé d'entrainer un algorithme d'apprentissage supervisé pour la reconnaissance de points caractéristiques (coins de la bouche, coins du nez…) en imagerie thermique qu'en imagerie visible. Citons à titre d'exemple les travaux de J. Ström basés sur des points non-spécifiques ainsi qu'un modèle 3D texturé du visage à *tracker* [66]. Enfin, précisions que ce type de méthode s'applique à n'importe quel objet tant qu'un modèle 3D texturé est disponible et que l'hypothèse d'objet rigide (c'est-à-dire non déformable) est respectée. L. Vachetti, V. Lepetit et P. Fua appliquent cette méthode à d'autres objets [67].

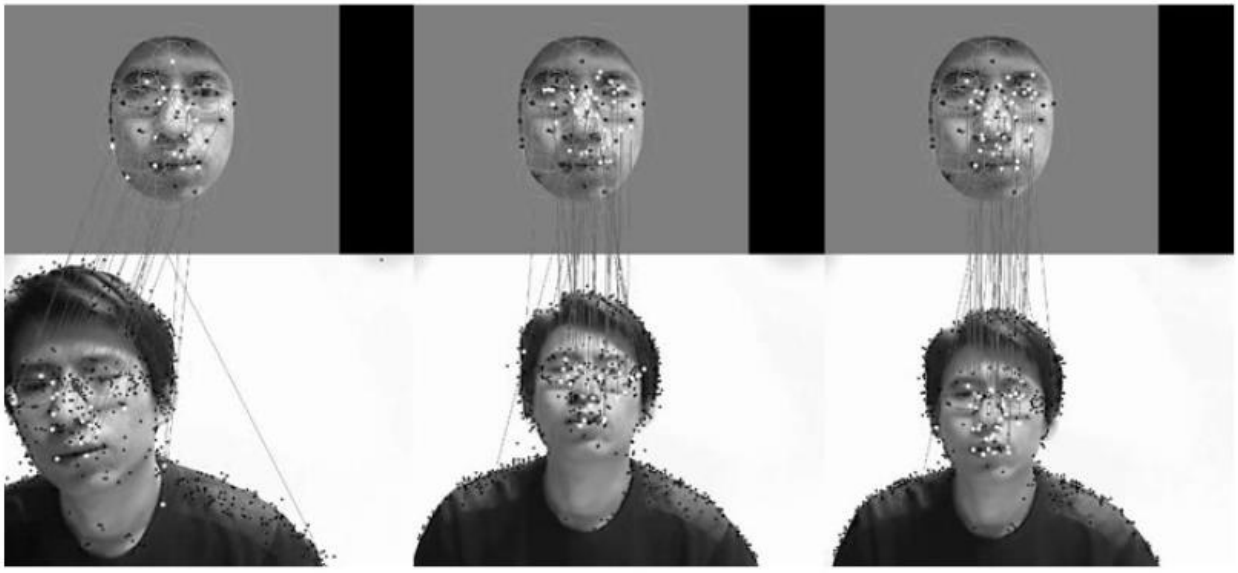


Figure 31. Méthode locale d'estimation de la pose fonctionnant avec des points d'intérêt non-spécifiques [65]. Sur chaque colonne on distingue une image de synthèse (première ligne) et une image réelle (deuxième ligne). Les segments reliant des points d'intérêt non-spécifiques entre une image de synthèse et une image réelle.

1.5.2. Les algorithmes d'estimation de la pose dans le *LWIR*

La détection par segmentation d'un visage humain semble facilitée en imagerie thermique grâce à la différence de niveau souvent marquée entre le visage et le fond. Pour illustrer cela traçons l'histogramme d'une image thermique acquise dans nos locaux (cf. Figure 32). On observe sur l'histogramme (cf. Figure 32 b)) plusieurs gaussiennes dont l'une représente les pixels de la peau. Dans les conditions du laboratoire, la peau est souvent l'élément le plus chaud et il est pertinent de considérer qu'une simple segmentation par seuillage est un bon point de départ pour détecter le visage.

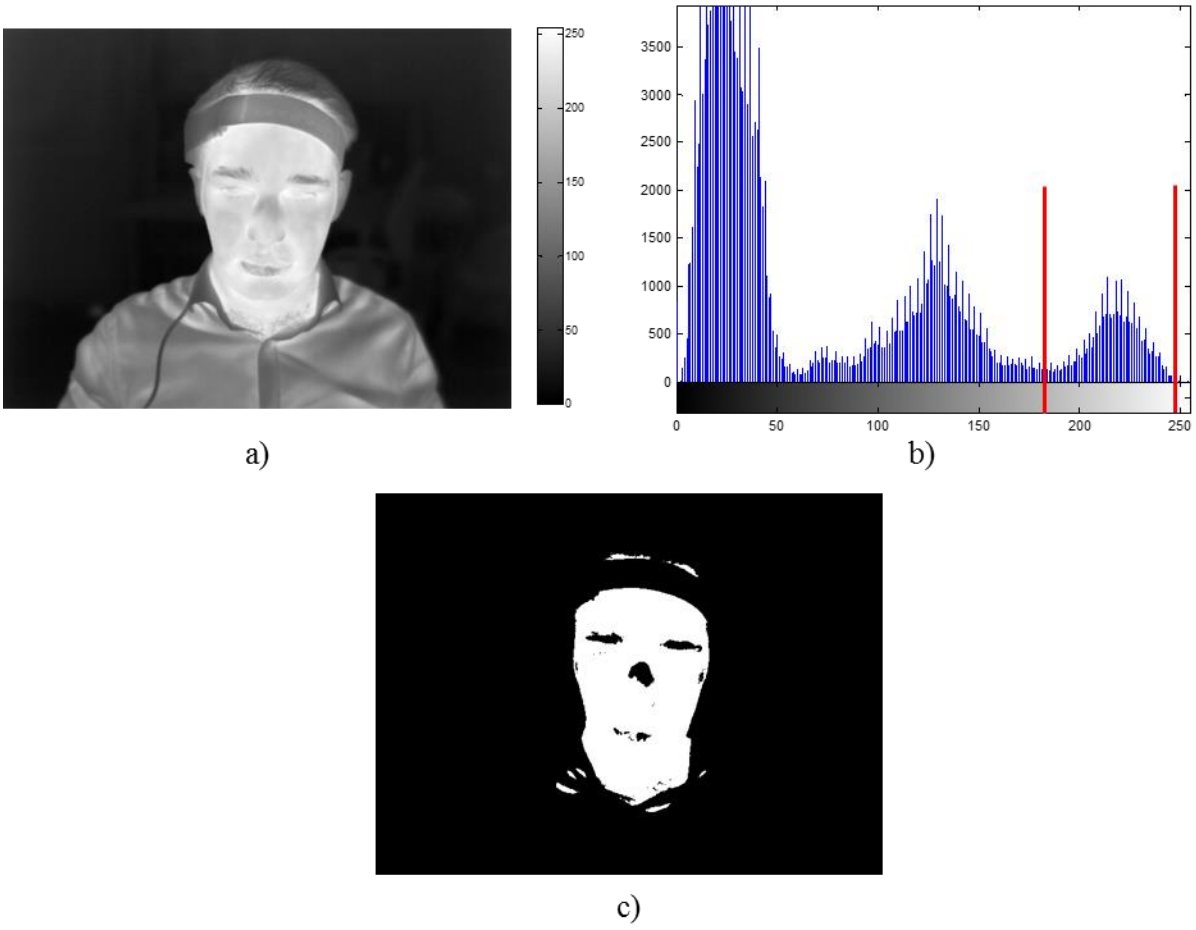


Figure 32. Segmentation du visage : a) image thermique d'un buste humain, b) histogramme de l'image et c) segmentation par seuillage (les pixels dont le niveau est compris dans l'intervalle délimité par les segments rouges sur l'histogramme sont à 1, les autres pixels sont à 0).

Basé sur ce constat, la référence [68] propose d'utiliser le masque issu d'une étape de segmentation pour estimer l'angle *roll* rotation dans le plan). En faisant l'hypothèse que la forme d'un visage peut être approximée par une ellipse, alors, l'orientation de son grand axe correspond à l'angle *roll* (cf. Figure 33).

Pour chaque ligne de l'image (axe y_i sur la Figure 33), la coordonnée colonne (axe des x_i sur la Figure 33) du grand axe de l'ellipse est considérée comme étant la moyenne des contours du masque :

$$x_i = \frac{x_i^l + x_i^r}{2} \tag{1.3}$$

Avec x_i^l le contour gauche du visage à la $i^{ème}$ ligne et x_i^r le contour droit du visage à la $i^{ème}$ ligne. Comme x_i et y_i sont liées par une fonction affine. L'orientation de l'ellipse correspond à la pente de cette fonction affine. Afin d'être robuste aux problèmes de segmentation (cf. Figure 33 a)), les auteurs proposent d'utiliser une méthode *single linkage clustering*. Celle-ci permet d'estimer les couples (x_i, y_i) qui décrivent mal l'orientation du grand axe (*outliers*).

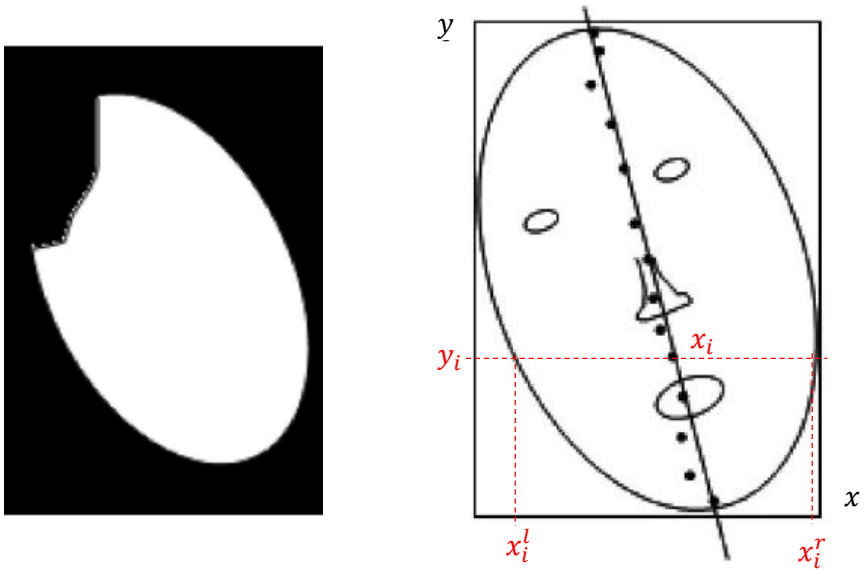


Figure 33. Estimation de l'angle roll d'un visage en imagerie thermique [68]. a) Simulation de masque imparfait du visage issu de la segmentation. b) estimation de l'angle *roll*. Il correspond à l'orientation du grand axe de l'ellipse.

Les auteurs s'appuient également sur la possibilité de segmenter le corps humain du fond en se basant sur une différence de niveau de gris de l'image thermique [69]. Les auteurs précisent que l'histogramme d'une image contenant un buste humain contient plusieurs gaussiennes dont l'une d'elles représente le buste. Toutes les gaussiennes sont testées et celle qui permet d'obtenir une forme qui ressemble à un buste est conservée. La tête est extraite du buste en utilisant l'hypothèse que la largeur du masque du buste est la plus fine au niveau du cou. A ce stade, le visage est localisé. Pour estimer son orientation, un algorithme d'analyse en composantes principales (*PCA* pour *principal component analysis*) entrainé sur une base d'images labellisées est implémenté. Lorsque que la *PCA* est utilisée dans le domaine de la reconnaissance de visage ou de l'estimation d'orientation, on parle de méthode *eigenfaces*.

Les auteurs [70] estiment que la gaussienne de l'histogramme, dont le centre est le plus élevé, correspond à la peau (on fait donc l'hypothèse que la peau est l'élément le plus chaud de l'image). Une fois que le visage est segmenté, les auteurs proposent de détecter une zone contenant le nez en faisant une hypothèse supplémentaire : le nez est l'élément le plus froid du visage. Ensuite en analysant la position du nez par rapport au barycentre de la zone du visage, l'angle *yaw* est estimé.

A notre connaissance, aucune méthode d'estimation de la pose (ou au moins de l'orientation) en imagerie thermique n'atteint les performances des méthodes implémentées en visible. Nous avons donc développé au cours de ma thèse nos propres algorithmes d'estimation de la pose en ayant pour objectif d'identifier la précision atteignable.

1.6. Synthèse des avantages et inconvénients d'un imageur *LWIR*

L'avantage le plus important de l'imagerie thermique est son invariance aux conditions d'illuminations ambiantes. Ainsi un simple détecteur passif peut fonctionner de nuit comme de jour et dans toutes les conditions d'illuminations difficiles. Un corollaire important de l'imagerie thermique pour les applications qui nécessitent l'analyse du visage des individus est la garantie du respect de la sécurité oculaire des sujets quel que soit la durée d'utilisation dudit système.

Cependant, les images thermiques sont connues pour avoir moins de texture que les images visibles [71].

Un autre inconvénient concerne les variations d'apparences thermiques inter-individus qui ne permettent pas intuitivement d'envisager l'utilisation d'algorithmes basés sur un apprentissage supervisé pour la détection de points spécifiques (coins du nez, de la bouche…), comme cela est souvent effectué dans le visible. Une recherche sur les algorithmes d'estimation de la pose adaptés à l'imagerie thermique nous est donc apparue être une étape essentielle à construire pour progresser sur le sujet de la surveillance du conducteur par imagerie thermique.

Concernant les applications de surveillance du conducteur, il est évident que l'estimation fine de la direction du regard ne pourra pas être adressée par l'imagerie *LWIR* car il est difficile de différencier les pupilles du cristallin. De plus, les verres de lunettes absorbent le rayonnement thermique. Par contre l'estimation de la direction de la tête semble tout à fait adressable par cette technologie. On rappelle que l'estimation de la direction de la tête approxime très souvent la direction du regard notamment dans les cas de rotation importante du visage et lorsque l'on souhaite connaître l'intention du conducteur.

L'estimation des mouvements de paupières semble inenvisageable en imagerie thermique car il est difficile de discerner les paupières du reste de l'œil. Cepednant, l'imagerie *LWIR* ouvre des perspectives intéressantes pour des applications de surveillances sans contact de paramètres physiologiques comme la fatigue, le stress ou la charge mentale, le rythme respiratoire, l'alcoolémie. Des travaux proposent même d'estimer le rythme cardiaque avec des caméras de haute qualité. La surveillance du conducteur, que cela soit son niveau d'attention ou de somnolence est une tâche complexe. Lorsqu'une tâche complexe est à réaliser automatiquement, la tendance naturelle est de multiplier le nombre de capteurs. Nous pensons donc

qu'une surveillance efficace du conducteur passera par la collecte puis l'analyse des informations produites par un jeu de plusieurs capteurs.

Chapitre 2. Etalonnage d'une caméra thermique

2.1. Introduction

Ce chapitre a pour objectif de présenter le matériel utilisé dans ce projet, c'est-à-dire la caméra thermique commerciale *Gobi 640 CL* développée par la société *Xenics*, un corps noir et une chambre climatique. Les notations associées aux images et aux différentes températures mises en jeu seront introduites (cf. section 2.2).

La *responsivité* (anglicisme provenant du terme *responsivity*, couramment utilisé par les fabricants de détecteurs) d'une caméra thermique en V/W ou en *Adu/W* sera définie (*Adu* pour *analog to digital output*). Il s'agit de la capacité d'un détecteur à convertir un rayonnement incident en une tension en volt. Nous évaluerons celle de la caméra commerciale *Gobi 640 CL* (cf. section 2.3).

Nous aborderons ensuite le bruit qui limite la technologie des caméras thermiques basées sur des matrices de microbolométres : le bruit temporel (section 2.4). Puis nous nous attarderons sur la correction du bruit spatial en imagerie thermique qui est encore un sujet de recherche actif (cf. section 2.5).

Enfin nous testerons une méthode d'étalonnage radiométrique (cf. section 2.6). Une image *LWIR* étalonnée ne dépend que du rayonnement de la scène. Cela peut être avantageux pour les algorithmes d'estimation de la pose, ce qui se traduit par une accélération du temps de calcul. Il est donc intéressant d'analyser la faisabilité d'un étalonnage radiométrique dans un habitacle automobile.

2.2. Matériel et notations

2.2.1. La caméra thermique (non-refroidie)

Nous utilisons une caméra *Gobi 640 CL*. Elle intègre une matrice de détecteurs microbolomètres développée par la société *Ulis* référencée *UL 04 322 640×480 17um.* Comme cela a déjà été évoqué à la section 1.2.2 un pixel bolomètre est sensible à la température. Afin qu'un pixel détecte en priorité le rayonnement de la scène, il est isolé de son environnement thermique grâce à des clous de suspension très fins (cf. Figure 34) et grâce à une mise sous vide (pour éviter la conduction thermique de l'air). L'absorption est maximisée dans le *LWIR*, grâce entre autre, à un réflecteur situé sous la membrane absorbante (cf. Figure 34). Ces deux éléments forment une cavité Fabry-Pérot dite $\lambda/4$: le déphasage entre l'onde incidente sur la membrane et l'onde réfléchie vaut $\lambda/4$, ce qui provoque l'absorption du rayonnement incident et réfléchi (on parle d'interférences destructives dans le cas de l'interférométrie).

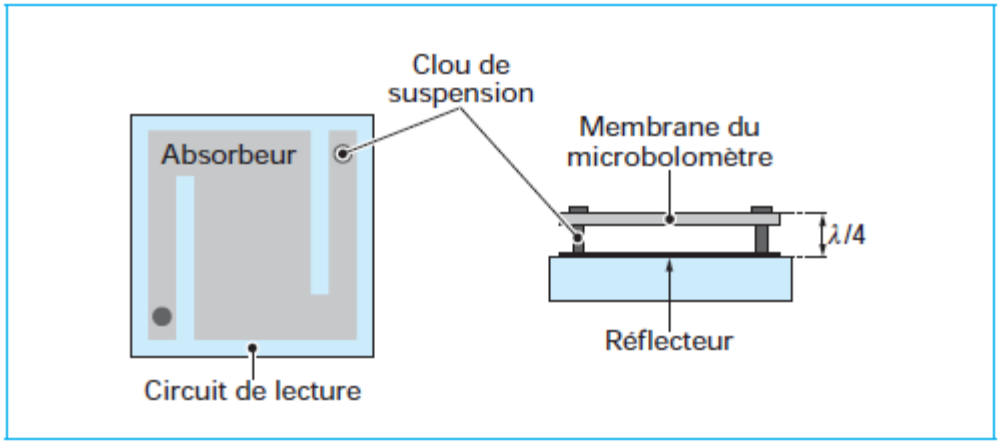


Figure 34. Principe de fonctionnement d'un pixel bolomètre.

Le boîtier de la caméra, l'optique et l'électronique rayonnent sur les pixels. Malgré le soin porté sur l'isolation thermique des pixels, ceux-ci peuvent également varier en température par conduction thermique. Finalement les fluctuations de température de tous ces éléments vont modifier l'image (cf. Figure 35). La caméra *Gobi 640 CL* ne possède que deux thermomètres. Le premier est fixé au boîtier et nous noterons la mesure de celui-ci T_C. Le second est fixé au plan focal et nous noterons la température qu'il mesure T_{FPA} (*FPA* pour *focal plane array*). La température T_{FPA} n'est pas accessible sur le *SDK* (*software development kit*) fournie par la société *Xenics*. Nous nous contenterons donc de la température du boîtier T_C et nous considérerons que la caméra est dans un état thermique où $T_{FPA} = T_C$. Notre objectif n'étant pas d'effectuer des mesures radiométriques très précises, cette approximation ne sera pas gênante.

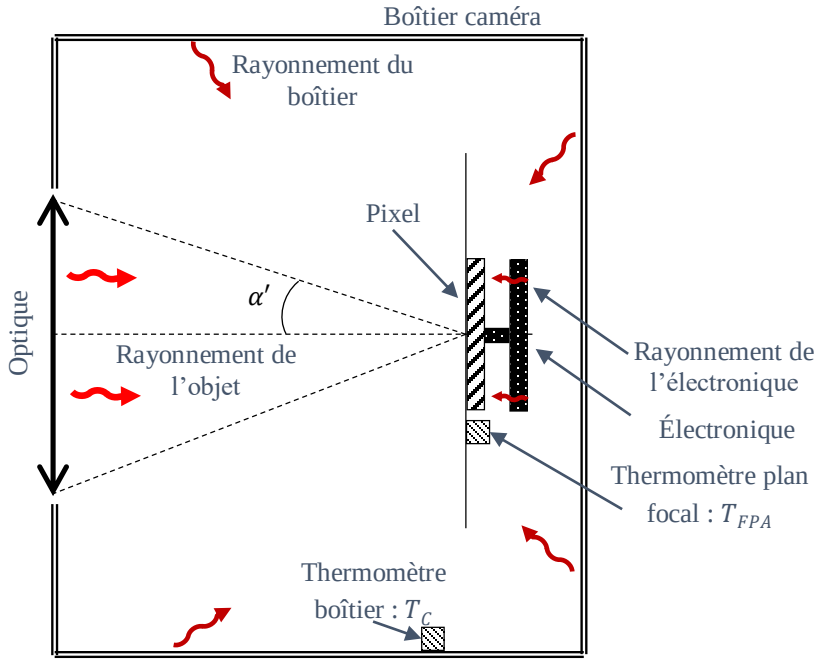


Figure 35. Représentation schématique des différents éléments qui rayonnent sur le pixel d'une caméra thermique non-refroidie.

Un circuit de lecture typique de pixels microbolomètres développé par la société *Ulis* est illustré par la Figure 36. Celui-ci est généralement adapté aux spécifications techniques propres à chaque fabricant, mais le principe reste toujours basé sur la même idée. Le signal de fond (température de l'optique, du boîtier, des pixels) est capté par des pixels aveugles et, il est ensuite soustrait au signal utile reçu par les pixels actifs [72]. La lecture en courant des pixels est effectuée grâce à des transistors alimentés avec des tensions de référence très peu bruitées [73]. La tension de référence des pixels actifs est notée *GFID* et celle des pixels aveugles est notée *GSK*. Ces tensions sont pilotables par l'utilisateur. Elles impactent l'offset et la dynamique du détecteur [74]. Le signal utile en courant est ensuite converti en tension grâce à un étage intégrateur *CTIA* (*capacitance trans-impedance amplifier*). L'intégration dans le temps permet de réduire le bruit. La capacité de charge et le temps d'intégration sont également pilotables par l'utilisateur. La réponse en tension est finalement numérisée grâce à un convertisseur analogique numérique *CAN*. Le signal à la sortie du *CAN* est numérisé sur 16 bits en niveaux que nous nommerons *Adu*

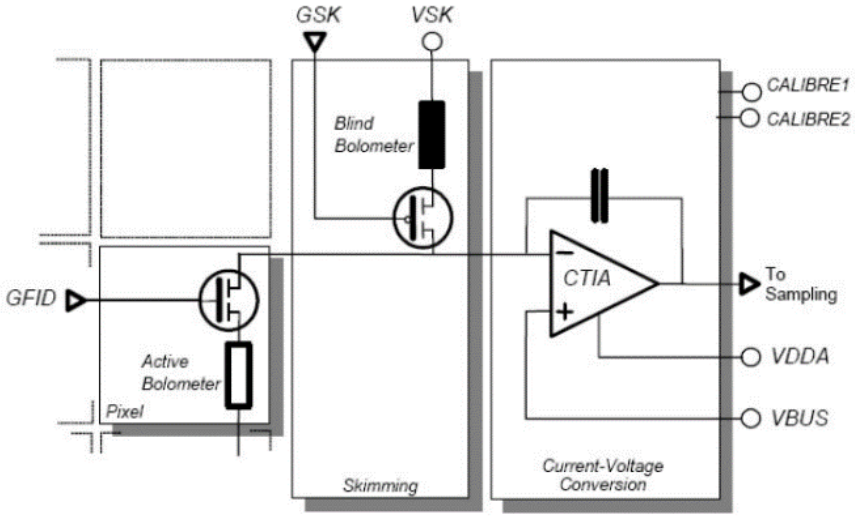


Figure 36. Circuit de lecture typique des matrices de microbolomètres développé par la société Ulis.

Nous avons utilisé l'ensemble des paramètres récapitulés dans le Tableau 1 qui sont des paramètres proches des réglages par défaut proposés sur la caméra *Gobi 640 CL*. Certains paramètres optiques et électroniques sont également rappelés dans la deuxième partie du Tableau 1.

Tableau 1. Paramétrage de la matrice de microbolomètres (et autres paramètres optiques et électroniques).

Paramètres réglables	**Valeurs**
GFID	1.5 V
GSK	2.969 V
Capacité de charge (*CTIA*)	6 pF
Temps intégration (*CTIA*)	25 µs
Paramètres fixes	**Valeurs**
Conversion analogique numérique	16 bits
Nombre d'ouverture optique f/#	0.8
Ouverture numérique (image)	0.53
Pas pixel	17 µm
Facteur de remplissage	>80 %

Précisons les différentes notations utilisées pour désigner les images de la caméra thermique. Les caractères en gras dans les formules signifient que l'on parle d'un vecteur ou d'une matrice. Une image d'une scène quelconque est représentée par une matrice $\boldsymbol{Y}$ de taille 480×640 car la caméra utilisée possède un format *VGA* (*Video Graphic Array*).

Remarque : les fabricants de détecteurs donnent généralement la largeur et ensuite la hauteur d'une matrice de pixels. Ainsi un format VGA serait plutôt 640×480. Dans le langage de programmation Matlab utilisé dans ces travaux de thèse, la première dimension d'une matrice correspond aux lignes (donc à la hauteur) et la seconde aux colonnes (donc à la largeur). C'est pourquoi nous indiquons que $\boldsymbol{Y}$ *possède une taille de 480×640.*

Introduisons dès à présent les notations utilisées pour la température. Un objet à température T_{objet} émet un rayonnement propre qui est '*vu*' par la caméra. Ce rayonnement est considéré comme la partie utile du signal total détecté par la caméra. Pour exprimer l'idée que l'objet n'est pas homogène en température, on définit la matrice $\boldsymbol{T}_{objet}$ où chaque élément représente la zone projetée d'un pixel du microbolomètre dans le plan objet (cf. Figure 37).

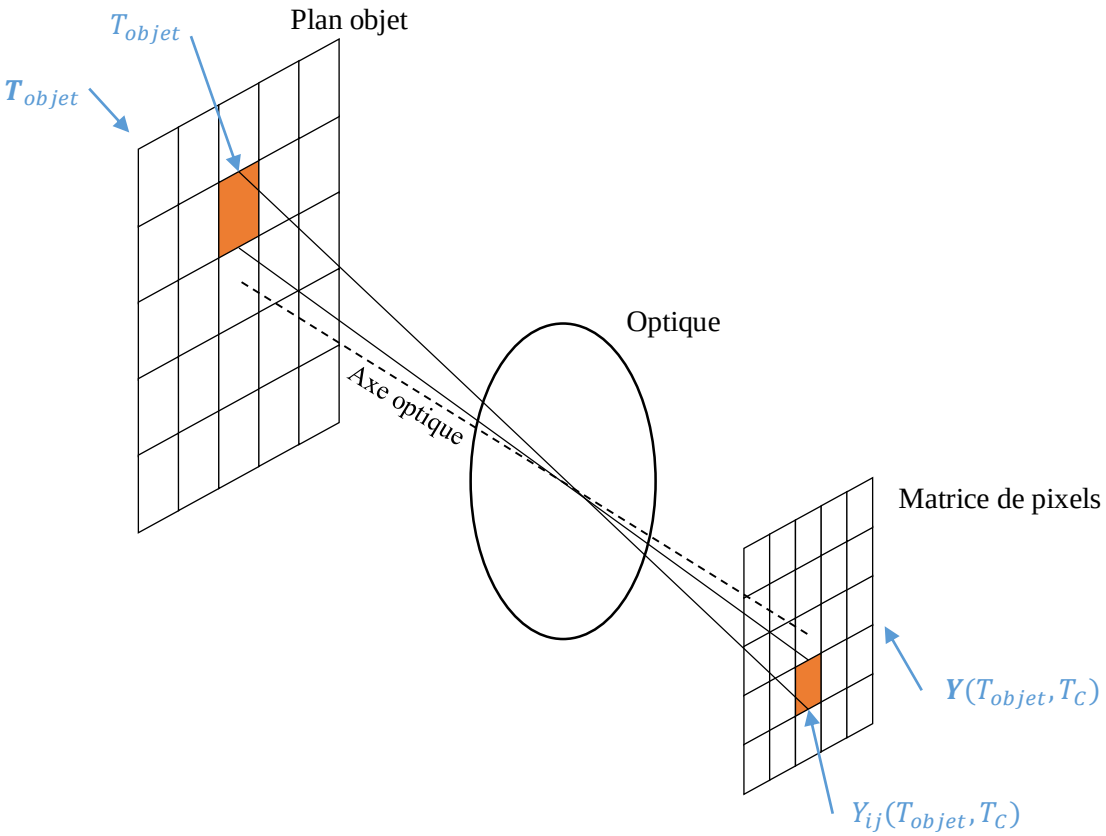


Figure 37. Représentation du système de notation au niveau du plan objet et de la matrice de pixels.

Comme déjà évoqué, la température de la caméra est notée T_C. Si on souhaite exprimer une image $\boldsymbol{Y}$ en apportant des précisions sur la température de l'objet et de la caméra, on indique en premier la température de l'objet et en second la température de la caméra comme suit : $\boldsymbol{Y}(\boldsymbol{T}_{objet}, T_C)$. La matrice $\boldsymbol{T}_{objet}$ a la même taille que la matrice $\boldsymbol{Y}$. La convention utilisée pour localiser les pixels est illustrée sur la Figure 38. L'origine est située sur le coin supérieur gauche de l'image. La valeur du pixel i, j de l'image (en Adu) pour une température objet $T_{objet}(i, j)$ et pour une température de caméra T_C est notée $Y_{ij}(T_{objet}(i, j), T_C)$.

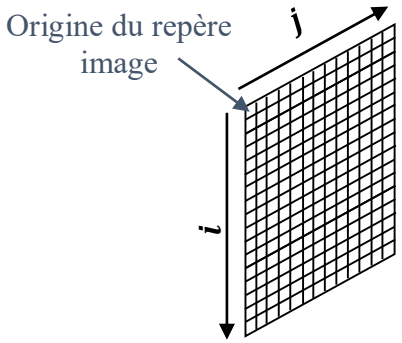


Figure 38. Repère image. Les lignes sont repérées par l'indice $i \in [1,480]$, les colonnes par l'indice $j \in [1,640]$.

Plusieurs travaux proposent de modéliser la réponse des pixels en fonction des différentes températures mises en jeu [75,76,77]. La précision de la modélisation est difficile à obtenir et ces travaux servent avant tout à aider à la conception de nouveaux circuits électroniques. Une caractérisation expérimentale du détecteur est toujours nécessaire. Pour caractériser et étalonner une caméra thermique on peut utiliser un corps noir plan. Cet objet étalon est présenté dans la section suivante.

2.2.2. Le corps noir plan

Un corps noir plan est un appareil électronique étalon qui possède une surface utile présentant des caractéristiques proches de celles d'un corps noir idéal : source Lambertienne, émissivité proche de l'unité. La température de la surface utile notée T_{CN} est mesurée avec précision et est pilotable électroniquement. Son uniformité spatiale et temporelle est également garantie. Un corps noir est utile pour évaluer la réponse d'une caméra thermique à une excitation connue.

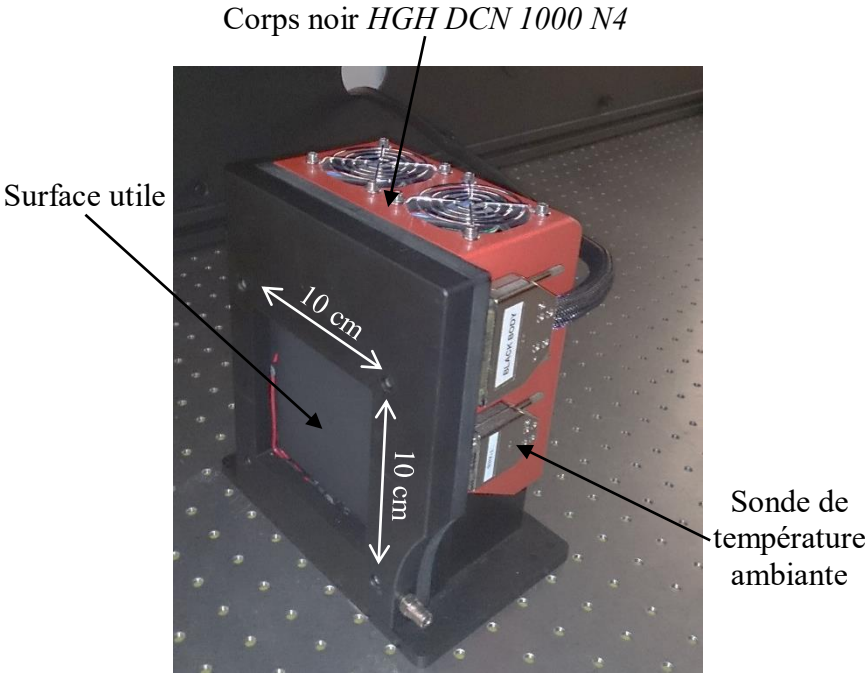


Figure 39. Corps noir plan *HGH DCN 1000 N4*.

Nous nous sommes procuré un corps noir *HGH DCN 1000 N4*. La surface utile est un carré de surface 10×10 cm² (cf. Figure 39). L'émissivité de la surface utile vaut $\varepsilon_{CN} = 0.98 \pm 0.02$. Sa température est pilotable dans un intervalle $T_{CN} \in [-5°C; +100°C]$. Deux sondes PT100 mesurent la température de la surface utile. La résolution du corps noir vaut 1 mK et est dictée par la résolution de ces deux sondes PT100. La précision sur la température absolue est définie par l'étalonnage usine réalisé chez *HGH*. En prenant en compte les différentes sources d'erreur, le fournisseur annonce une incertitude à ±2σ autour de plusieurs valeurs nominales données dans le Tableau 2.

Tableau 2. Incertitude sur la valeur nominale (c'est-à-dire la valeur de consigne) du corps noir *HGH DCN 1000 N4*.

Valeurs nominales (°C)	0	20	50	80	150
Incertitude ±2σ (°C)	1.1	0.70	0.50	0.40	0.40

L'image du corps noir obtenue grâce à la caméra thermique est notée $\boldsymbol{O}$. Au premier ordre, cette image dépend de la température du corps noir et de celle de la caméra, on notera cette image $\boldsymbol{O}(T_{CN}, T_C)$. La valeur du pixel i, j de l'image (en Adu) à la température du corps noir T_{CN} et pour une température de caméra T_C est notée $O_{ij}(T_{CN}, T_C)$.

Lorsque que l'on acquiert plusieurs images d'un corps noir successivement dans le temps pour un couple de températures (T_{CN}, T_C), on utilise la notion de cube notée $\boldsymbol{C}(T_{CN}, T_C)$ (cf. Figure 40). Un cube est une matrice de taille $480 \times 640 \times K$. La lettre K représente le nombre d'images acquises successivement au cours du temps. Nous avons utilisé $K = 50$. La matrice $\boldsymbol{C}_k(T_{CN}, T_C)$ de taille 480×640 correspond à une image du corps noir. La moyenne temporelle d'un cube (c'est-à-dire le long de sa troisième dimension) est notée $\overline{\boldsymbol{C}}(T_{CN}, T_C)$.

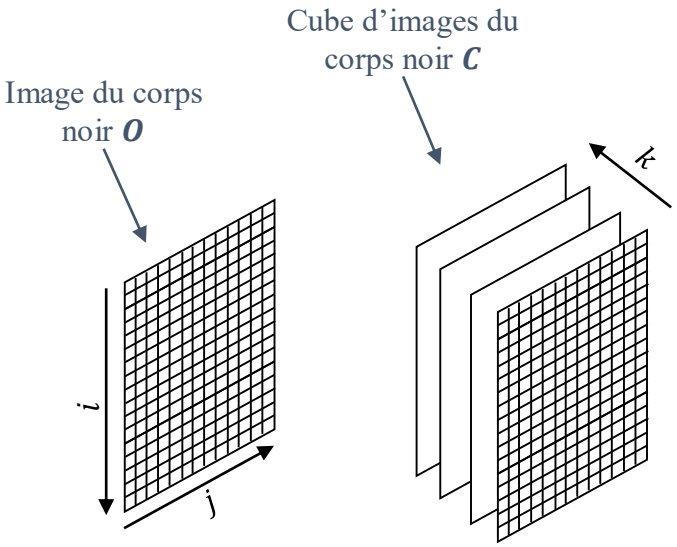


Figure 40. Définition d'un cube d'images. Les lignes sont notées i et les colonnes j. Ce sont des coordonnées spatiales 2D. La succession des images est repérée par l'indice k.

2.2.3. La chambre climatique

Pour connaître la réponse de la caméra à un couple de températures (T_{CN}, T_C) on utilise une chambre (ou enceinte) climatique et un corps noir qui recouvre tout le champ de vue *FOV* (*field of view*) de la caméra (cf. Figure 41). La température T_{CL} de la chambre climatique est pilotable et modifie la température de la caméra T_C.

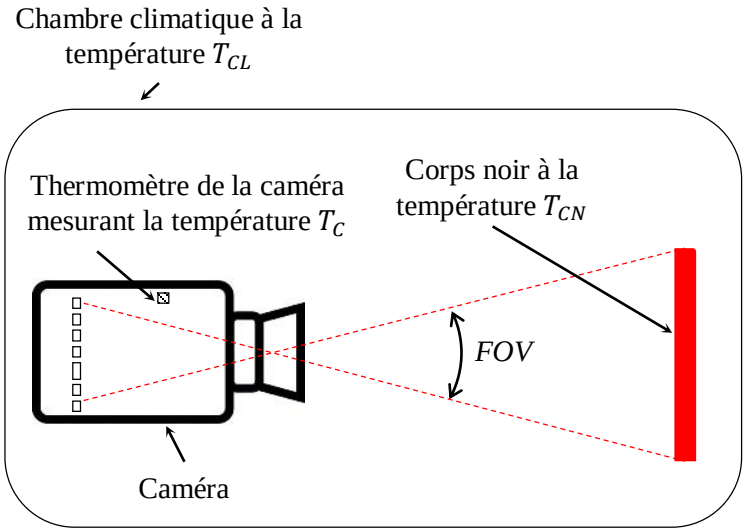


Figure 41. Schéma d'illustration des différentes températures mises en jeu en imagerie thermique non refroidie.

Nous avons utilisé une chambre climatique *CLIMATS 2223TE* appartenant au service des essais thermiques de Renault (cf. Figure 42).

Le dernier rapport de vérification réalisé selon la norme NFX 15.140 à la date du 13/04/2016 mentionne des valeurs de stabilité de la température mesurées grâce à 9 sondes PT100 certifiées placées à 9 positions données dans la chambre climatique (la chambre climatique est alors vide). La stabilité de la température correspond à la variation maximale pendant la durée de la mesure (>60 minutes). Les valeurs de stabilité mesurées à partir de la sonde PT100 placée au centre de la chambre climatique sont répertoriées dans le Tableau 3.

Tableau 3. Stabilité de la température dans la chambre climatique à la position approximative de la caméra thermique.

Température moyenne (en °C)	**Stabilité (en °C)**
-40.86	0.92
64.31	0.62
150.33	0.82

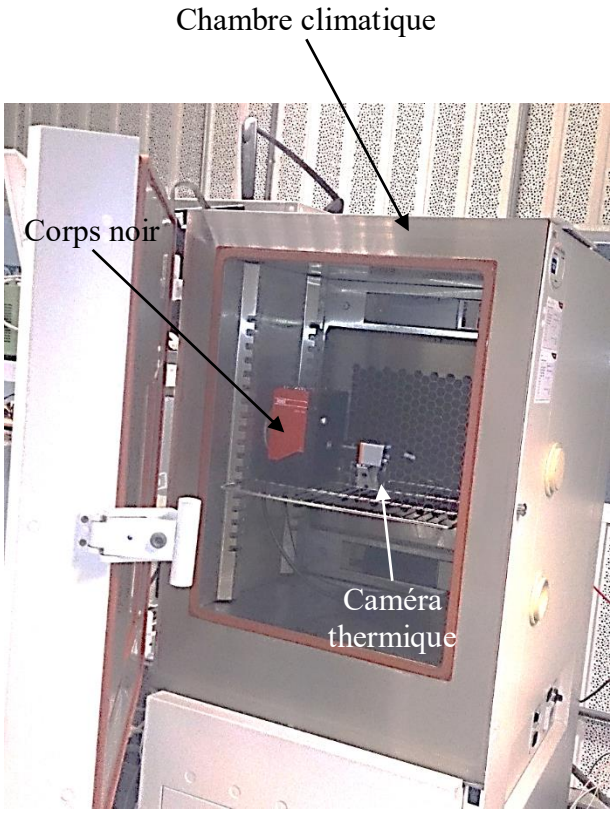
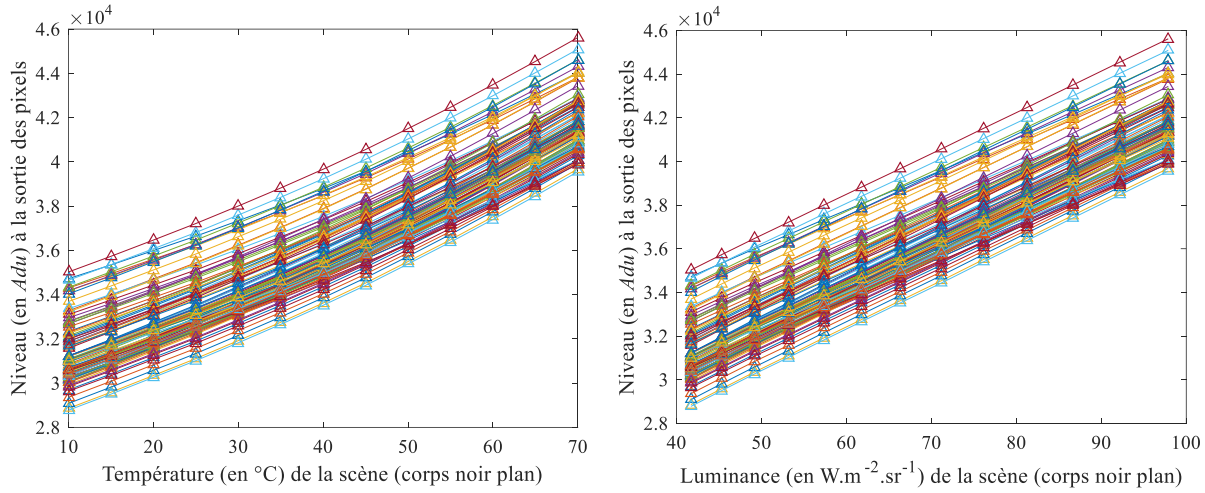


Figure 42. Photographie d'un banc de caractérisation d'une caméra thermique composé d'une chambre climatique, d'un corps noir et d'une caméra thermique non-refroidie.

2.3. *Responsivité* en V/W (ou en *Adu*/W)

La réponse d'une caméra (ou *responsivité,* terme adapté de l'anglais *responsivity*) mesure la capacité d'un détecteur à convertir un rayonnement incident en une tension en volt. Dans notre cas, l'entrée est un flux reçu par un pixel (en W) et la sortie est une tension (en V) qui est convertie en *Adu* (*analog to digital output*) par un *CAN* (Convertisseur Analogique Numérique).

Un détecteur thermique possède une réponse non linéaire à la température de la scène observée, mais linéaire au rayonnement incident (cf. Figure 43). La différence vient de la loi de Planck (cf. Annexe A, équation (A.1)) qui lie la température de l'objet à son rayonnement de manière non-linéaire. Il est ainsi plus aisé de caractériser la réponse au rayonnement qu'à la température de la scène car seulement deux points de mesure sont nécessaires (cependant sur une petite plage de température, la réponse à la température peut être considérée comme linéaire).



a) b)

Figure 43. Réponses brutes (c'est-à-dire sans correction des non-uniformités spatiales) des pixels de la caméra *Gobi 640 CL* à un corps noir plan de température variable (on acquiert 50 images du corps noir et on réalise la moyenne temporelle). Cette série de mesures a été réalisée hors d'une chambre climatique, à température ambiante. La température de la caméra varie dans un intervalle de 0.42°C tout au long des acquisitions. a) Réponse en température de la scène. b) Réponse au rayonnement de la scène.

Pour connaitre la *responsivité*, on utilise un objet dont le rayonnement est connu (le corps noir) et on compare le signal résultant du détecteur à deux températures différentes de corps noirs, c'est-à-dire à deux quantités de flux différentes. La *responsivité* des pixels de la caméra en (Adu/W) est définie comme suit :

$$\mathfrak{R} = \frac{\overline{\boldsymbol{C}}(T_{CN2}, T_C) - \overline{\boldsymbol{C}}(T_{CN1}, T_C)}{\boldsymbol{F}_{scène}(T_{CN2}) - \boldsymbol{F}_{scène}(T_{CN1})} \tag{2.1}$$

La *responsivité* $\mathfrak{R}$, le flux rayonné par la scène $\boldsymbol{F}_{scène}$ et la réponse du détecteur (en V ou en *Adu*) $\overline{\boldsymbol{C}}$ sont des matrices de la même taille que la matrice de pixel (480×640), car chaque pixel possède sa propre réponse.

La réponse $\mathfrak{R}$ dépend de la température de caméra T_C. Celle-ci doit-être constante au moment de l'acquisition des cubes d'images $\boldsymbol{C}(T_{CN1}, T_C)$ et $\boldsymbol{C}(T_{CN2}, T_C)$. De plus, on utilise la moyenne temporelle d'un cube plutôt qu'une seule image d'un corps noir (c'est-à-dire qu'on utilise $\overline{\boldsymbol{C}}(T_{CN}, T_C)$ au lieu de $\boldsymbol{O}(T_{CN}, T_C)$) afin de réduire la contribution du bruit temporel dans l'estimation de la réponse.

Pour évaluer $\boldsymbol{F}_{scène}(T_{CN})$, on intègre la loi de Planck dans la bande passante du détecteur (cf. Annexe A, équation (A.2)), ce qui nous donne la radiance (ou luminance) de la source, puis on la multiplie par l'étendue géométrique qui dépend des paramètres optiques de la caméra. Il est également nécessaire de prendre en compte la transmission de l'atmosphère, la transmission de l'optique, l'absorption des pixels, etc (cf. Annexe A, équation (A.3))…

Il est également possible de définir la responsivité en radiance (dont l'unité est en $V.W^{-1}.m^2.sr$ ou en $Adu.W^{-1}.m^2.sr$). Celle-ci, notée $\mathfrak{R}'$ présente l'avantage d'être linéaire et ne nécessite pas la connaissance des paramètres optiques du système. C'est donc cette responsivité que nous utiliserons plus volontiers dans ce chapitre.

Nous avons évalué la *responsivité* $\mathfrak{R}'$ de la caméra *Gobi 640 CL* à différentes températures T_C de la caméra. Pour cela nous plaçons la caméra face à un corps noir dans une enceinte climatique. Le champ de vue *FOV* de la caméra est totalement recouvert par le corps noir. Nous faisons 13 points de mesure de la réponse $\mathfrak{R}$: la température de la chambre climatique varie de -5°C à +55°C par pas de 5°C. Pour chaque point de mesure, nous attendons 1 heure pour que la température de la caméra soit *stable*, c'est-à-dire $\frac{\partial T_C}{\partial t} = 0$. Puis nous enregistrons 50 images brutes du corps noir à 30°C et 50 images brutes du corps noir à 40°C (cf. Figure 44).

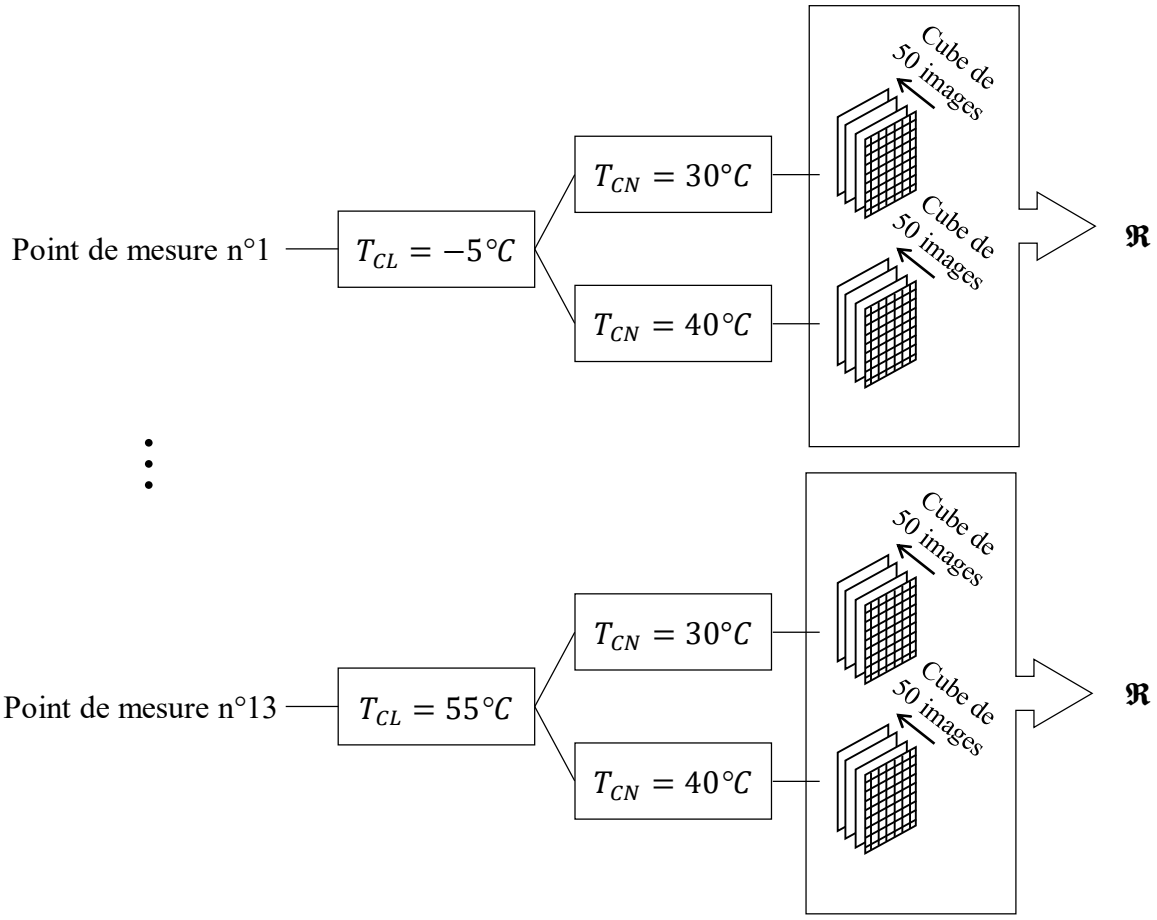


Figure 44. Procédure pour la mesure de la responsivité à plusieurs températures ambiantes T_{CL}.

Le Tableau 4 répertorie les moyennes des températures de la caméra T_C pour chaque cube : il s'agit d'une moyenne temporelle sur les 50 images qui forment un cube.

Tableau 4. Température T_C de la caméra pour les différents points de mesure de la responsivité. Toutes les températures sont en dégrés Celsius.

Température de la chambre climatique (°C)	**Le corps noir est à 30°C**	**Le corps noir est à 40°C**	**Différence ΔT_C (en °C) de la température de la caméra entre l'acquisition du cube à 30°C et à 40°C**
	Moyenne de la température de la caméra T_C sur les 50 images d'un cube (en °C)		
-5	8.70	8.53	0.17
0	13.28	13.22	0.06
5	17.95	17.99	-0.04
10	22.69	22.76	-0.07
15	27.52	27.55	-0.03
20	32.32	32.37	-0.04
25	37.25	37.30	-0.05
30	42.19	42.23	-0.04
35	47.14	47.17	-0.03
40	52.21	52.23	-0.03
45	57.39	57.49	-0.10
50	62.28	62.32	-0.04
55	68.28	68.32	-0.04

La différence de température de la caméra ΔT_C entre l'acquisition du cube d'images lorsque le corps noir est à 30°C et lorsque le corps noir est à 40°C limite la précision de l'estimation de la *responsivité* (cf. quatrième colonne du Tableau 4). Cette différence est en partie due à une incertitude sur la régulation de la température de la chambre climatique.

Les estimations des réponses de 12 pixels de la matrice de bolomètres (choisis aléatoirement) à différentes températures de la caméra sont illustrées sur la Figure 45.

Pour évaluer le flux rayonné par le corps noir, nous avons calculé sa luminance spectrique L_λ grâce à la loi de Planck pour la température de consigne du corps noir, et nous avons intégré L_λ entre 8 et 14 µm.

Le bruit spatial a été corrigé grâce à une correction basée sur un obturateur mécanique (cf. section 2.5.5).

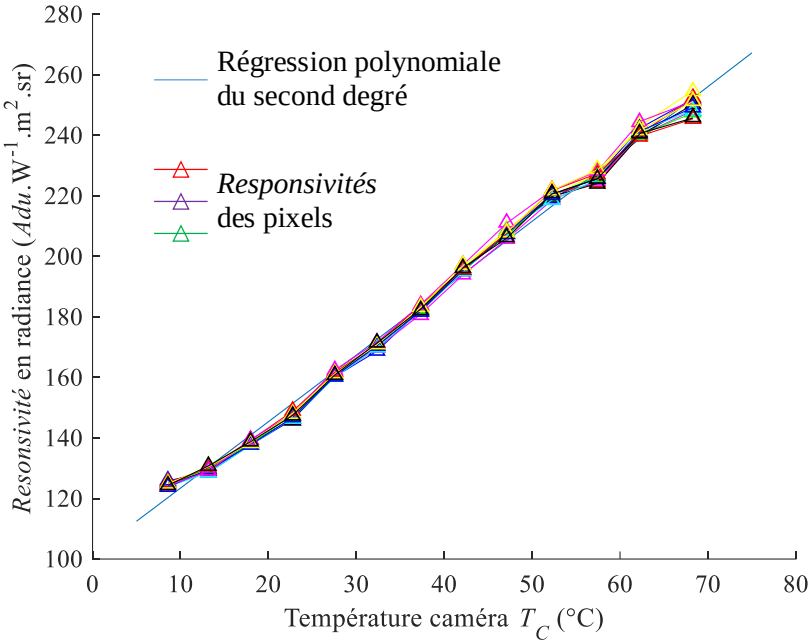


Figure 45. Estimation de la *responsivité* (en *Adu*/W^{-1}.m².sr) de 12 pixels de la caméra *Gobi 640 CL* en fonction de la température du boîtier T_C. La *responsivité* est estimée à partir de deux cubes d'images d'un corps noir à 30 et à 40°C. Les images sont corrigées du bruit spatial fixe grâce à un obturateur mécanique. La courbe en trait plein bleu est la régression polynomiale de la moyenne spatiale des responsivités des 640×480 pixels.

Nous réalisons une régression polynomiale d'ordre 2 de la moyenne spatiale des réponses en fonction de la température de la caméra (cf. courbe en trait plein bleu sur la Figure 45). Puis nous conservons les trois tables de coefficients $\boldsymbol{A}_{Resp}$, $\boldsymbol{B}_{Resp}$ et $\boldsymbol{C}_{Resp}$ qui permettront de calculer la réponse en fonction de la température de la caméra :

$$\mathfrak{R}' = \boldsymbol{A}_{Resp}(T_C)^2 + \boldsymbol{B}_{Resp}T_C + \boldsymbol{C}_{Resp} \tag{2.2}$$

2.4. Bruit temporel

Le bruit temporel s'évalue pour chaque pixel. Pour l'estimer, on place la caméra face à un corps noir et on enregistre un cube d'images $\boldsymbol{C}(T_{CN}, T_C)$. Ensuite, l'écart type selon la dimension temporelle du cube est calculé :

$$\boldsymbol{\sigma}_{\boldsymbol{temporel}} = \sqrt{\frac{\sum_{k=1}^{K}\left(\boldsymbol{C}_{\boldsymbol{k}}(T_{CN}, T_C) - \overline{\boldsymbol{C}}(T_{CN}, T_C)\right)^2}{K}} \tag{2.3}$$

La matrice $\boldsymbol{\sigma}_{\boldsymbol{temporel}}$ représente le bruit temporel de chaque pixel. Elle peut être exprimée en volt ou en *Adu*. Le bruit temporel équivalent en température est souvent utilisé pour spécifier la sensibilité d'une caméra thermique. Le terme *NETD* (*noise equivalent temperature difference*) désigne cet équivalent. La

NETD correspond à la différence de température entre un objet et le fond qui produit un rapport signal sur bruit qui vaut 1.

$$NETD = \frac{\tilde{v}}{\Delta V}\Delta T \tag{2.4}$$

Le bruit en volt est noté $\tilde{v}$. La différence de température ΔT correspond à la différence de température entre le fond et l'objet d'intérêt. Enfin ΔV correspond à la différence de réponse en volt entre le fond et l'objet d'intérêt. On peut également exprimer la relation (2.4) en unité *Adu* :

$$NETD = \frac{\sigma_{temporel}}{\Delta U}\Delta T \tag{2.5}$$

La différence des réponses (en *Adu*) au fond et à l'objet d'intérêt est notée ΔU.

Pour évaluer la *NETD*, il y a plusieurs méthodes comme cela est expliqué dans la référence [78] aux pages 117-120. L'annexe B présente une déduction de la *NETD* à partir de la mesure du bruit en *Adu*, de la responsivité en *Adu*/W et des paramètres opto-mécaniques. Nous utiliserons ici une méthode expérimentale qui fait intervenir la *responsivitié* en *Adu*/K. La *responsivité* en *Adu*/W est linéaire. Par contre la réponse en *Adu*/K ne l'est pas. Cependant sur un petit intervalle de température, il est possible de la considérer linéaire. Dans ce cas, la *NETD* peut directement s'exprimer ainsi :

$$\boldsymbol{NETD} = \frac{\boldsymbol{\sigma}_{temporel}}{\boldsymbol{\Re}_{Adu/K}} \tag{2.6}$$

En nous inspirant d'une méthode d'estimation de la *NETD* d'un fournisseur de caméra [79], nous avons procédé à une mesure de la *NETD* de la caméra *Gobi640 CL*. Nous avons évalué la *NETD* à une température objet qui nous intéresse, c'est-à-dire celle de la peau humaine qui est proche de 34°C.

On acquiert trois séries de données dans des conditions identiques, c'est-à-dire que la température du laboratoire, la température du plan focal et température du boîtier (etc…) sont stables :

- un cube d'images du corps noir à 30°C,
- un cube d'images du corps noir à 34°C,
- un cube d'images du corps noir à 40°C.

La sonde de température du boîtier de la caméra vaut approximativement $T_C = 38°C$ pour ces trois acquisitions. Les cubes d'images à 30°C et 40°C servent à calculer la matrice des réponses en *Adu*/K notée $\mathfrak{R}_{Adu/K}$. Le cube d'images à 34°C sert à calculer la matrice des écarts types $\sigma_{temporel}$. On représente généralement les écarts types grâce à un histogramme (cf. Figure 46). On obtient ensuite la *NETD* grâce à l'équation (2.6). On représente également la *NETD* à l'aide d'un histogramme (cf. Figure 47).

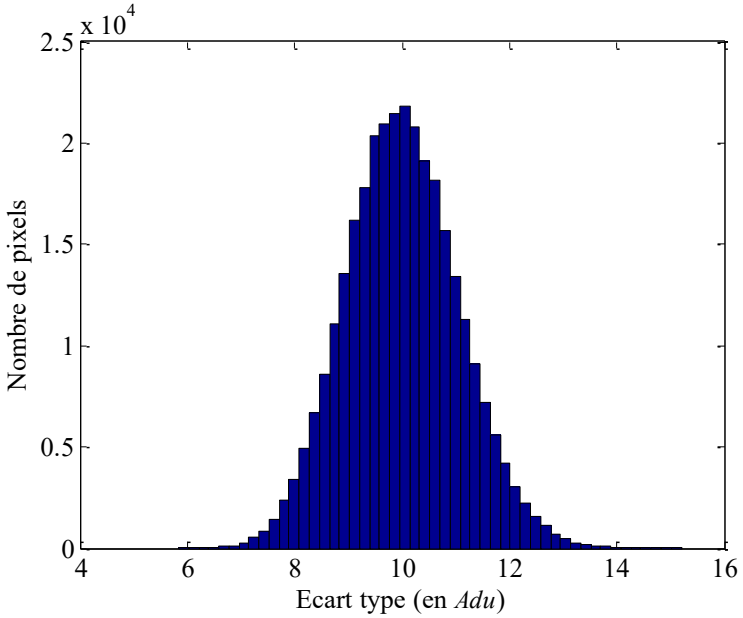


Figure 46. Bruit temporel en *Adu (analog to digital output).*

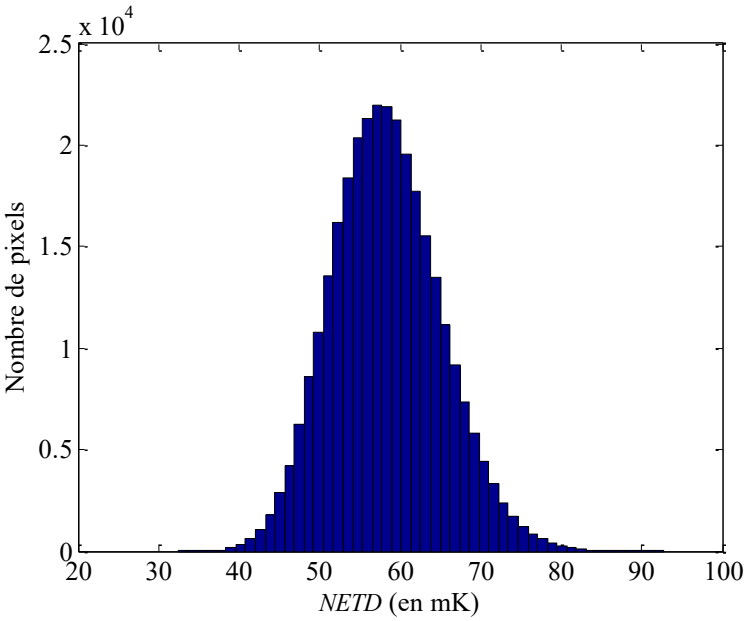


Figure 47. Histogramme de la *NETD* réalisé à partir d'un corps noir plan à 35°C pour une caméra à 38°C.

2.5. Bruit spatial fixe

2.5.1. Définition

Le bruit spatial fixe *BSF* désigne les disparités des réponses des détecteurs élémentaires Y_{ij} d'un plan focal éclairé par un fond uniforme généré par un corps noir pour une bande spectrale d'analyse imposée [80]. Le bruit spatial est souvent modélisé par une fonction affine [81,82]. L'image idéale $\boldsymbol{Z}$ d'un corps noir est affectée par un gain $\boldsymbol{g}$ et un offset $\boldsymbol{o}$.

$$\boldsymbol{Y} = \boldsymbol{g}.\boldsymbol{Z} + \boldsymbol{o} \tag{2.7}$$

La Figure 48 représente l'image d'un corps noir. Pour quantifier le *BSF* on réalise généralement l'écart type spatial de l'image d'un corps noir que l'on note $\boldsymbol{\sigma}_{BSF}$ et que nous exprimons en *Adu* (il est également souvent exprimé en volt) :

$$\boldsymbol{\sigma}_{BSF} = \sqrt{\frac{\sum_{i=1}^{480}\sum_{j=1}^{640}\left(O_{ij}(T_{CN},T_C) - \langle \boldsymbol{O}(T_{CN},T_C)\rangle\right)^2}{480 * 640}} \tag{2.8}$$

La notation $\langle \boldsymbol{O}(T_{CN},T_C)\rangle$ désigne la moyenne spatiale de l'image $\boldsymbol{O}(T_{CN},T_C)$.

Pour atténuer la dispersion des réponses on établit des coefficients. On peut par exemple utiliser une correction type « deux points » (cf. section 2.5.2). Après correction, le *BSF* est réduit, le bruit restant est appelé bruit spatial fixe résiduel (*BSFR*) [80].

Les origines du *BSF* sont multiples. Sur la Figure 48, on distingue en premier lieu un effet de colonnage. Celui-ci est relatif à la conception du circuit électronique de lecture. En effet, les pixels d'une colonne possèdent des éléments électroniques communs. On distingue également sur la Figure 48 un bruit de basse fréquence spatiale. Celui-ci a plusieurs origines dont l'atténuation de l'éclairement du centre de la matrice vers les bords (due à l'optique), la variation spatiale de la température du plan focal, la variation spatiale du rayonnement thermique du boîtier de la caméra... Enfin, il existe un *BSF* à l'échelle du pixel de type poivre et sel qui provient de la dispersion technologique des microbolomètres. Ce dernier apparaît plus clairement lorsque le bruit de colonne est corrigé.

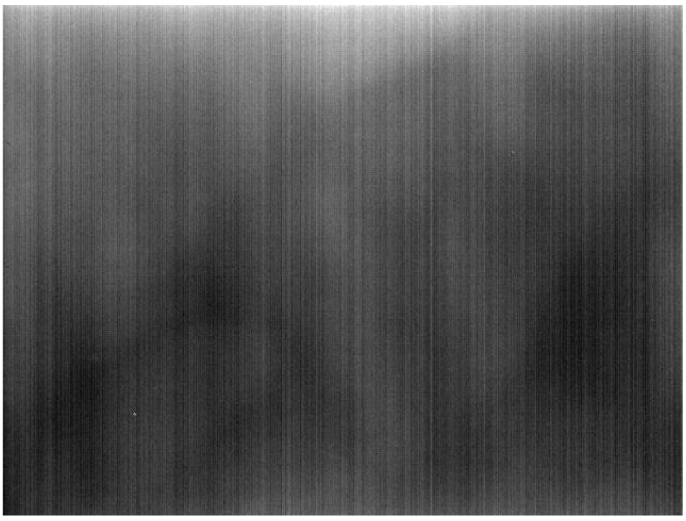


Figure 48. Carte du *BSF* de la caméra *Xenics Gobi 640 CL*. Il s'agit de l'image d'un corps noir à la température $T_{CN} = 30°C$ pour une température de caméra $T_C = 8.7°C$.

2.5.2. Correction « deux points »

Le terme qui désigne la correction du *BSF* est *NUC* pour *Non-Uniformity-Correction*. Une correction « deux points » est un type de *NUC* simple et souvent utilisé [83]. Le terme « deux points » exprime l'idée que dans l'équation (2.7), il y a deux inconnues : le gain $\boldsymbol{g}$ et l'offset $\boldsymbol{o}$. Ces deux points de mesure seront représentés par la moyenne temporelle de deux cubes de 50 images du corps noir à deux températures T_{Froid} et T_{Chaud}. Ces deux points de mesure sont notés $\boldsymbol{I}_{Froid}$ et $\boldsymbol{I}_{Chaud}$. La correction de l'image $\boldsymbol{Y}$ s'exprime comme suit :

$$\boldsymbol{Y_{corrigée}} = (\boldsymbol{Y} - \boldsymbol{I}_{Froid})\frac{\langle \boldsymbol{I}_{Chaud}\rangle - \langle \boldsymbol{I}_{Froid}\rangle}{\boldsymbol{I}_{Chaud} - \boldsymbol{I}_{Froid}} + \langle \boldsymbol{I}_{Froid}\frac{\langle \boldsymbol{I}_{Chaud}\rangle - \langle \boldsymbol{I}_{Froid}\rangle}{\boldsymbol{I}_{Chaud} - \boldsymbol{I}_{Froid}}\rangle \tag{2.9}$$

Réécrivons l'équation (2.9) pour faire apparaître un offset de correction et un gain de correction

$$\boldsymbol{Y_{corrigée}} = \boldsymbol{A}.\boldsymbol{Y} + \boldsymbol{B}, \tag{2.10}$$

avec

$$\boldsymbol{A} = \frac{\langle \boldsymbol{I}_{Chaud}\rangle - \langle \boldsymbol{I}_{Froid}\rangle}{\boldsymbol{I}_{Chaud} - \boldsymbol{I}_{Froid}}, \tag{2.11}$$

et

$$\boldsymbol{B} = -\boldsymbol{I}_{Froid}\boldsymbol{A} + \langle \boldsymbol{I}_{Froid}\boldsymbol{A}\rangle. \tag{2.12}$$

Remarque : le terme $\langle \boldsymbol{I}_{Froid}A\rangle$ de l'offset de correction $\boldsymbol{B}$ permet de ne pas impacter le niveau continu de l'image suite à une correction 2 points.

Une correction « deux points » est exacte qu'aux deux températures de la scène T_{Chaud} et T_{Froid}. Entre ces deux températures, et à l'extérieure le *BSFR* augmente en imagerie refroidie. En imagerie thermique (non-refroidie), lorsque les deux températures T_{Chaud} et T_{Froid} sont proches, on n'observe pas

d'augmentation du *BSFR* entre ces deux températures. Par contre à l'extérieur, on observe bien une augmentation du *BSFR*.

Le choix d'une température chaude et d'une température froide permet d'ajuster l'intervalle de température dans lequel l'image sera la mieux corrigée. Nous choisissons d'opter pour une température chaude de 40°C et une température froide de 30°C car nous souhaitons obtenir la meilleure qualité d'image pour la peau du visage dont la température est comprise entre 32°C et 34°C [84].

$$\begin{aligned} \boldsymbol{I}_{Chaud} &= \overline{\boldsymbol{C}}(T_{CN} = 40°C, T_C) \\ \boldsymbol{I}_{Froid} &= \overline{\boldsymbol{C}}(T_{CN} = 30°C, T_C) \end{aligned} \tag{2.13}$$

La qualité d'une *NUC* est déterminée par le bruit spatial fixe résiduel *BSFR*. On évalue le *BSFR* grâce à l'écart type spatial $\boldsymbol{\sigma}_{BSFR}$ de l'image après la correction *NUC*.

On calcule donc $\boldsymbol{\sigma}_{BSFR}$ sur la caméra *Gobi 640 CL* après une *NUC* « deux points » dont les points de référence sont représentés par (2.13) (cf. Figure 49). On évalue $\boldsymbol{\sigma}_{BSFR}$ à plusieurs températures du corps noir T_{CN} et à température de caméra T_C fixe.

En imagerie, la correction est estimée suffisante lorsque le *BSFR* est de l'ordre de grandeur du bruit temporel. Sur la Figure 49, le bruit temporel a donc été ajouté à titre de comparaison. On constate que dans un intervalle approximatif de 20°C à 45°C, après correction *NUC* « deux points » le *BSFR* est inférieur au bruit temporel. Le choix des températures chaude et froide est donc bien adapté à notre application. La Figure 50 illustre qualitativement l'effet d'une correction « deux points » des non-uniformités.

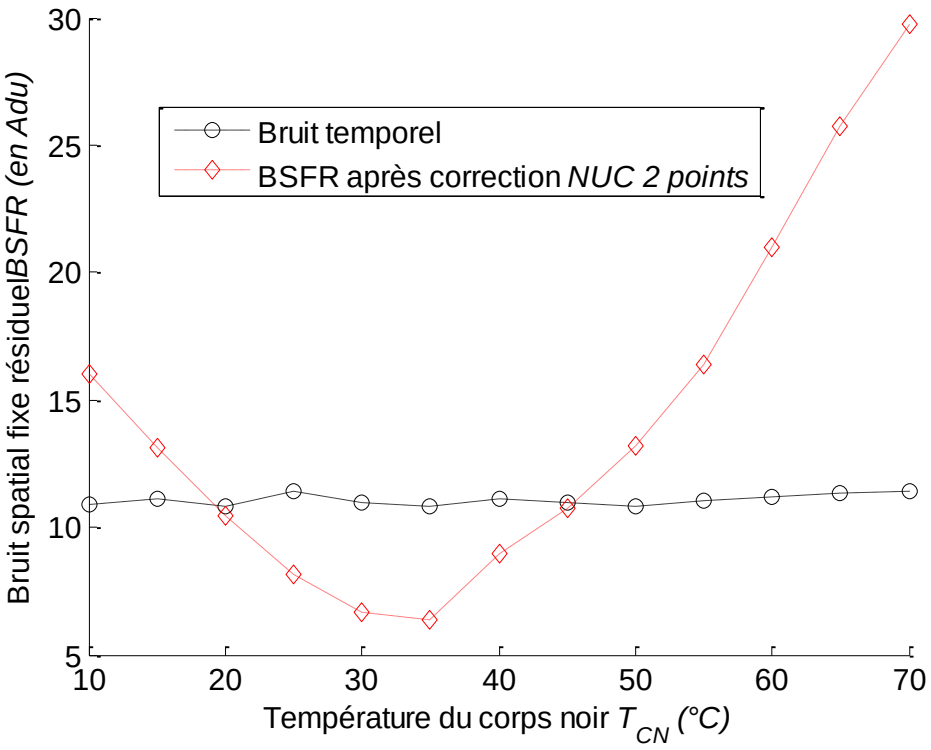


Figure 49. Représentation du *BSFR* après correction *NUC* « deux points » (courbe rouge en pointillés) et du bruit temporel (ligne discontinue noire). La correction « deux points » a été établie grâce à une température froide de 30°C et une température chaude de 40°C.

a) b)

Figure 50. Illustration de l'effet d'une correction NUC « deux points ». L'image brute a) et l'image corrigée b) d'un même visage sont représentées. La correction « deux points » a été établie grâce à une température froide de 30°C et une température chaude de 40°C.

2.5.3. Dépendance thermique de la correction « deux points »

Lorsque la température T_C de la caméra évolue, il est nécessaire de rafraîchir la *NUC*. Le *BSF* est donc dépendant de la température du capteur. (Rappel : on fait l'hypothèse que la température T_C de la caméra est égale à la température T_{FPA} du plan focal).

Dans la littérature il est souvent signalé que l'offset de correction $\boldsymbol{B}$ évolue avec la température de la caméra T_C alors que le gain de correction $\boldsymbol{A}$ y est relativement indépendant [81]. Nous l'avons également constaté en faisant l'expérience suivante. La caméra *Gobi 640 CL* est positionnée face à un corps noir dans une chambre climatique. Pour une température de consigne de la chambre climatique T_{CL}, on acquiert deux cubes d'images du corps noir, l'un à la température froide ($T_{Froid} = 30°C$), l'autre à la température chaude ($T_{Chaud} = 40°C$) et on calcule la table de gain $\boldsymbol{A}$ et la table d'offset $\boldsymbol{B}$. On répète cette opération pour d'autres températures de la chambre climatique T_{CL}. Un délai d'une heure à chaque température de consigne de la chambre climatique est écoulé avant l'acquisition des cubes pour établir le *NUC* « deux points ». La caméra est donc dans un état thermique relativement stable à chaque acquisition.

La Figure 51 illustre l'offset de correction pour un ensemble donné de pixels en fonction de la température de la caméra T_C. La Figure 52 illustre le gain de correction pour le même ensemble de pixels en fonction de la température de la caméra T_C.

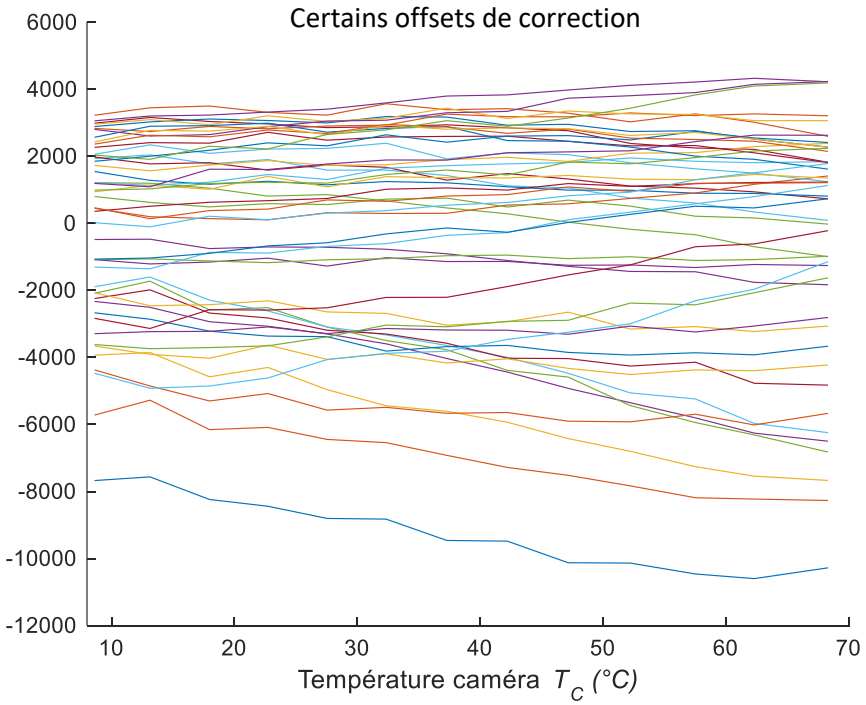


Figure 51. Offset de correction de certains pixels en fonction de la température de la caméra T_C.

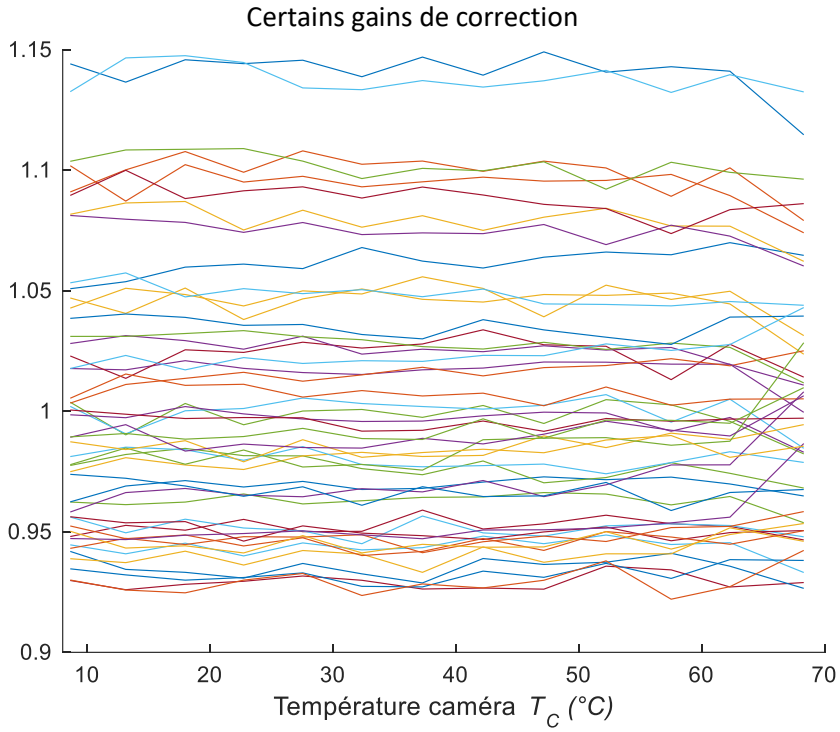


Figure 52. Gain de correction de certains pixels en fonction de la température de la caméra T_C.

On constate effectivement que l'offset est dépendant de la température de la caméra T_C alors que le gain y est relativement indépendant. Il est envisageable d'établir une table de gains de correction dès la production de la caméra et de la conserver tout au long de la vie de cette dernière. Nous appelons cette table, table de gains de correction usine $\boldsymbol{A}_{usine}$. Pour l'établir, l'équation (2.11) montre qu'il nécessaire d'acquérir seulement deux cubes d'images du corps noir à deux températures T_{CN} pour une température de la caméra T_C fixe.

En ce qui concerne la table d'offsets de correction $\boldsymbol{B}$, il est indispensable de prévoir une méthode pour le mettre à jour en fonction de la température de la caméra T_C. Historiquement les fabricants évitaient cette complication en utilisant un élément de stabilisation en température (un Peltier par exemple). On parlait alors de *TEC* (*thermoelectric cooling*). Afin de proposer des produits compatibles avec les contraintes de coût et de *SWaP* (*size weight and power*) du marché civil et du marché militaire (le système doit être embarquable sur un fusil pour la vision de nuit des fantassins), les constructeurs ont supprimé le *TEC*, on parle alors de *TEC-Less*. En contrepartie, il est nécessaire d'ajouter un élément de référence pour rafraîchir la *NUC* lorsque la température de la caméra évolue. Une solution souvent mise en œuvre chez les constructeurs a été d'utiliser un obturateur mécanique ou *shutter*, qu'il est possible de fermer lorsqu'il est nécessaire de rafraîchir la *NUC*. Les fabricants ont cherché des solutions pour supprimer le *shutter* (on parle alors de système *shutterless*) pour deux raisons. La première est une raison de coût et de contraintes *SWaP*, la seconde est le fait que l'image est figée lorsque le *shutter* est fermé pour la mise à jour de la *NUC*. La méthode garantissant la meilleure qualité d'image en fonctionnement *shutterless* repose sur un étalonnage thermique de la caméra. L'étalonnage est cependant très long et par conséquent agit sur le coût final du

système. Plus récemment, la communauté scientifique et industrielle a proposé des méthodes *shutterless* sans recours à un étalonnage thermique. Ces méthodes sont basées sur des algorithmes de traitement d'images qui utilisent généralement l'*apriori* que l'image thermique est relativement lisse. Ces méthodes sont appelées *scene-based* (basées sur l'image de la scène).

Nous allons détailler les méthodes qui permettent de rafraîchir la *NUC* en commençant par les méthodes les plus performantes jusqu'aux méthodes, à l'heure actuelle, les moins performantes. Nous aborderons d'abord les méthode basées sur l'étalonnage thermique, puis les méthodes basées sur un *shutter* et enfin les méthodes basées sur les images de la scène (*scene-based*).

2.5.4. Correction *shutterless* basée sur un étalonnage

L'idée est de compenser la dérive de la *NUC* en se basant sur la température de la caméra T_C et un modèle du *BSF* (plutôt l'offset du *BSF* car on a vu que le gain était relativement indépendant de T_C). Pour créer le modèle du *BSF* des données sont acquises dans une phase d'étalonnage. Cette idée a été exploitée dans les références [81,82,85,86]. Nous faisons dans cette section une explication d'une implémentation particulière que nous avons réalisée et nous évaluons sa performance.

La table de correction du gain $\boldsymbol{A}_{usine}$ est évaluée une seule fois et nous la considérons indépendante de T_C. Nous enregistrons des tables d'offset de correction $\boldsymbol{B}$ pour un ensemble de température du détecteur T_C :

$$\boldsymbol{B} = -\boldsymbol{I}_{Froid}\boldsymbol{A}_{usine} + \langle \boldsymbol{I}_{Froid}\boldsymbol{A}_{usine} \rangle. \tag{2.14}$$

Puis, pour chacun des pixels, une régression polynomiale de degré p est réalisée. Ainsi $p+1$ tables de coefficients vont permettre d'estimer l'offset de correction $\widehat{\boldsymbol{B}}$ en fonction de la températures de la caméra :

$$\widehat{\boldsymbol{B}}(T_C) = \sum_{i=0}^{p} \boldsymbol{coef}_i \times (T_C)^i \tag{2.15}$$

Les termes $\boldsymbol{coef}_i$ sont des matrices 480×640. Le terme $(\boldsymbol{T_C})^i$ correspond à la température de la caméra à la puissance i. Nous établissons les tables de coefficients $\boldsymbol{coef}_i$ grâce à des acquisitions dans une chambre climatique dont la température T_{CL} varie de -5°C à +55°C par pas de 5°C.

Ensuite, pour tester la *NUC* « deux points » mise à jour par étalonnage, un autre ensemble de cube $\boldsymbol{C}(T_{CN} = 35°C, T_C)$ a été acquis. Dans ce dernier, la température du corps noir est toujours à $T_{CN} = 35°C$ et la température de la chambre climatique T_{CL} varie de -2.5°C à +52.5°C par pas de 5°C. On applique à ce nouveau jeu de données la correction « deux points » dont la table d'offset, régie par l'équation (2.15), utilise les coefficients $\boldsymbol{coef}_i$ qui ont été calculés à partir du premier jeu de mesures (T_{CL} variant de -5°C à +55°C par pas de 5°C). La température T_C a donc été enregistrée avec les images à corriger). Le *BSFR* est ensuite évalué (cf. Figure 53). Nous avons testé plusieurs degrés pour la régression polynomiale : second, troisième et quatrième degré. Le polynôme du second degré engendre un *BSFR* plus élevé que le bruit temporel, ce qui n'est pas satisfaisant. Nous choisissons le polynôme de degré 4 car c'est celui qui permet d'obtenir le *BSFR* le plus faible.

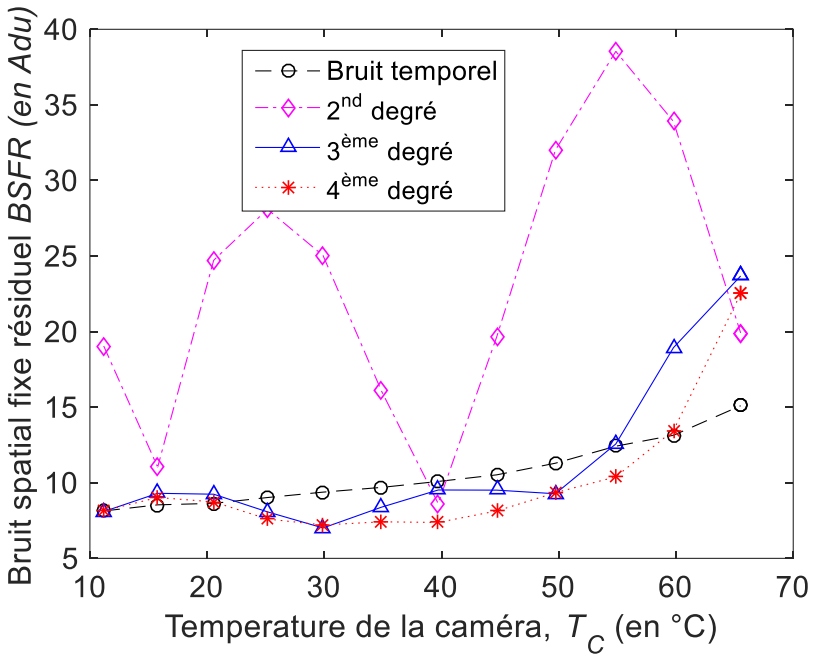


Figure 53. Bruit spatial fixe résiduel (*BSFR*) après une correction *NUC 2 points* basée sur l'étalonnage de la table d'offsets. Le bruit temporel est reporté sur le graphique à titre de comparaison (courbe noire avec des cercles). Le *BSFR* est mesuré après une correction obtenue grâce à un polynôme du second degré (courbe magenta avec des losanges), du troisième degré (courbe bleue avec des triangles), du quatrième degré (courbe rouge avec des $*$).

Précisons que cette méthode de correction de l'image, bien que très efficace, est très longue à réaliser. Nous devons attendre une heure pour que la température de la caméra soit stabilisée pour acquérir les images du corps noir. La durée d'acquisition des 13 points de mesure (la température de la chambre climatique T_{CL} va de -5°C à +55°C par pas de 5°C) est donc de 13 heures.

Pour mettre en évidence la nécessité d'attendre que la caméra soit stabilisée en température pour l'étalonnage, nous avons réalisé un autre test. La chambre climatique a été programmée pour qu'elle passe de 10°C à 70°C le plus rapidement possible. La température de la caméra T_C va donc évoluer tout au long de ce cycle. Les cubes $\boldsymbol{I}_{Froid}$ ont été acquis « à la volée », sans attendre que la température de la caméra soit stabilisée. On peut parler d'étalonnage *dynamique*. La Figure 54 représente la température de la caméra en fonction du temps. Les traits verticaux rouges représentent les points de mesures pour lesquels nous enregistrons un cube $\boldsymbol{I}_{Froid}$. Les traits obliques représentent les dérivés temporelles de la température par rapport au temps $\frac{\partial T_C}{\partial t}$. Cet étalonnage dure 70 minutes.

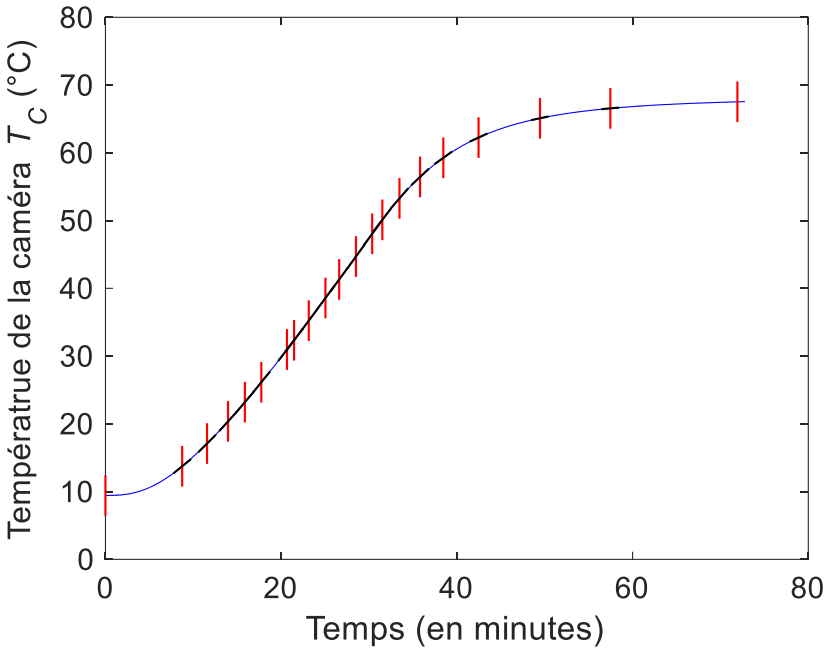


Figure 54. Température de la caméra en fonction du temps (courbe bleue) dans un essai d'étalonnage *dynamique* de la caméra. Les traits verticaux rouges représentent les points des mesures pour lesquels nous enregistrons l'offset de correction. Les traits en noir représentent les dérivés de la température en fonction du temps.

Nous utilisons une régression polynomiale de degré 4 pour établir l'offset en fonction de la température $\widehat{\boldsymbol{B}}$. Puis nous testons cette nouvelle *NUC* « deux points » sur l'ensemble de cube $\boldsymbol{C}(T_{CN} = 35°C, T_C)$ acquis pour des températures de la chambre climatique T_{CL} variant de -2.5°C à +52.5°C par pas de 5°C (cf. Figure 55). Cette expérience consiste donc à appliquer une correction acquise dans des conditions où $\frac{\partial T_C}{\partial t}$ est variable à des images acquises dans des conditions où $\frac{\partial T_C}{\partial t} = 0$. La Figure 54 et la Figure 55 montrent que plus $\frac{\partial T_C}{\partial t}$ est élevé plus le bruit spatial fixe résiduel *BSFR* est important.

Interprétation : les pixels bolomètres détectent le rayonnement thermique de tout leur environnement, c'est-à-dire le flux rayonné par la scène à travers l'optique, mais également l'intérieur du boîtier de la caméra. En régime dynamique, c'est-à-dire lorsque $\frac{\partial T_C}{\partial t} \neq 0$, il est difficile de connaitre les températures relatives des éléments composant la caméra car les constantes de temps thermiques sont différentes. Il est donc dangereux et inexact d'utiliser un seul thermomètre T_C comme référence de température. Les résultats illustrés sur la Figure 55 nous paraissent donc logiques.

Finalement nous constatons que l'étalonnage de l'offset est efficace s'il est réalisé dans un régime de température de la caméra *stable*, c'est-à-dire lorsque $\frac{\partial T_C}{\partial t} = 0$, et si la correction *NUC* « deux points » qui en découle est appliquée sur des images acquises également dans un régime de température de la caméra *stable*. Mais le temps d'étalonnage en régime de température de la caméra *stable* est long et on peut imaginer que le *BSFR* est élevé lorsque la caméra démarre car elle change rapidement de température.

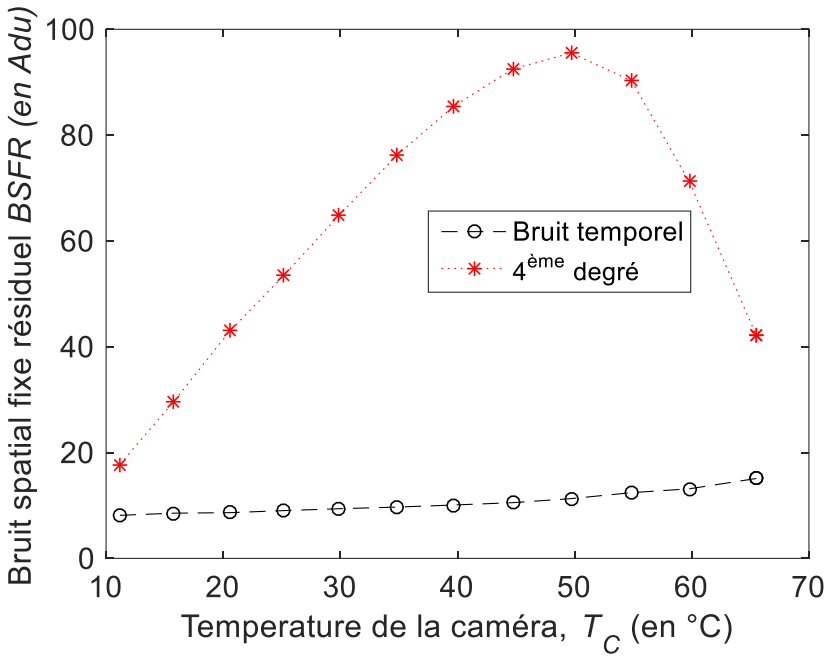


Figure 55. Bruit spatial fixe résiduel *BSFR* après *NUC 2 points dynamique* (courbe en pointillés rouge). Le bruit temporel est reporté également sur ce grahique à titre de compraison (courbe en trait discontinu noir).

Nous sommes conscients qu'une bonne connaissance de l'environnement thermique dans lequel est implantée la caméra pourrait permettre d'envisager une sorte d'étalonnage hybride :

- étalonnage de l'offset en régime *stable* pour certaines plages de température de la caméra,
- étalonnage de l'offset en régime *dynamique* pour d'autres plages de température de la caméra.

Des essais avec une caméra équipée de thermomètres supplémentaires (au moins sur l'optique, le boîtier et sur le plan focal) seraient nécessaires pour explorer la faisabilité d'une telle approche.

La *NUC* basée sur un *shutter* va être présentée dans la section suivante.

2.5.5. Correction *NUC* basée sur un *shutter* (obturateur mécanique)

Cette correction consiste à déclencher la fermeture d'un *shutter* lorsqu'il est nécessaire de rafraichir la *NUC*. La plupart des constructeurs proposent des caméras avec un *shutter* interne, c'est-à-dire un *shutter* placé entre l'optique et le plan focal. L'image du *shutter* $\boldsymbol{Y}_{Shutter}$ remplace l'image froide $\boldsymbol{I}_{Froid}$ dans la correction « deux points » :

$$\boldsymbol{Y_{corrigée}} = (\boldsymbol{Y} - \boldsymbol{Y}_{Shutter})\frac{\langle \boldsymbol{I}_{Chaud}\rangle - \langle \boldsymbol{I}_{Froid}\rangle}{\boldsymbol{I}_{Chaud} - \boldsymbol{I}_{Froid}} + \langle \boldsymbol{Y}_{Shutter}\frac{\langle \boldsymbol{I}_{Chaud}\rangle - \langle \boldsymbol{I}_{Froid}\rangle}{\boldsymbol{I}_{Chaud} - \boldsymbol{I}_{Froid}}\rangle \qquad (2.16)$$

La caméra *Xenics Gobi 640 CL* possède un *shutter* interne. A la mise en route de la caméra, le *shutter* est déclenché et une correction *NUC* « deux points » est établie grâce à l'équation (2.16). Lorsque la température de la caméra T_C évolue au-delà d'un certain seuil à spécifier par l'utilisateur, le *shutter* est à

nouveau déclenché et une nouvelle table de correction *NUC* « deux points » est établie. Par défaut, le *shutter* est déclenché pour des variations supérieures ou égales à 0.5°C.

Discussion sur la position de l'obturateur mécanique (le shutter) : le *shutter* est positionné entre l'optique et le plan focal. Il est n'est donc pas rigoureux d'appliquer un gain qui prend en compte l'atténuation d'éclairement dû à l'optique au terme $(\boldsymbol{Y} - \boldsymbol{Y}_{Shutter})$. Ainsi l'image corrigée par ce type de méthode présente un bruit basse fréquence (cf. Figure 56).

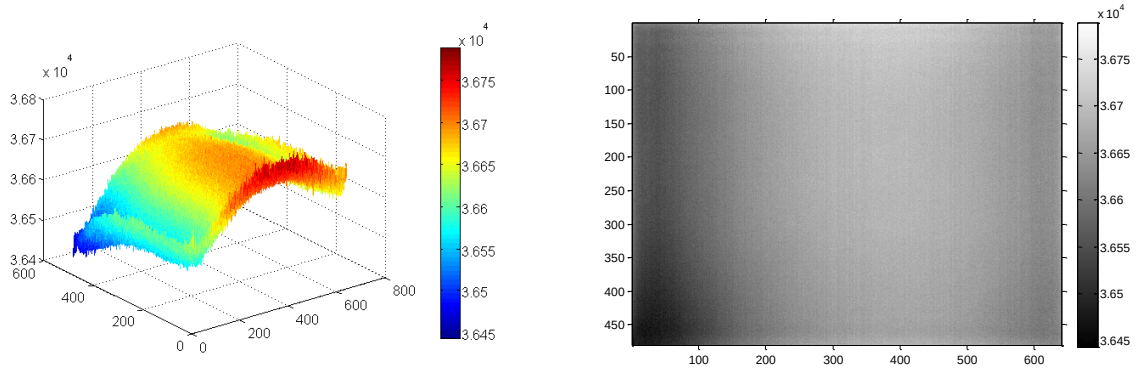


Figure 56. Image d'un corps noir à la température 30°C après une correction NUC 2 points basée sur un obturateur mécanique. Les niveaux de l'image sont exprimés en *Adu* (16 bits).

Dans ces conditions, le *BSFR* est élevé lorsqu'il est calculé sur la totalité de l'image (c'est-à-dire sur les 640×480 pixels). Cependant si on réduit la fenêtre de calcul du *BSFR*, on peut se rapprocher des valeurs obtenues avec une correction *NUC* par étalonnage (cf. Figure 58). On constate ainsi sur cette figure que le *BSFR* calculé sur une fenêtre de 20×20 pixels après une *NUC* basée sur le *shutter* atteint le *BSFR* calculé sur toute l'image (c'est-à-dire sur les 640×480 pixels) après une correction *NUC* basée sur un étalonnage.

Finalement la *NUC* basée sur un *shutter* est efficace à l'échelle d'une fenêtre de 20×20 pixels. Cette taille de fenêtre doit être gardée à l'esprit et être mise en relation avec les applications qui viennent ensuite. Nous montrerons dans le chapitre 5 qu'une méthode de l'état de l'art en traitement d'images utilise une zones de 16×16 pixels pour décrire un point d'intérêt utile à l'estimation de la pose 3D [87].

Remarque : pour être tout à fait rigoureux des zones plus grandes peuvent également être utilisées pour décrire les points d'intérêt. La taille de cette zone est déterminée automatiquement dans la méthode expliquée dans la réf [87]. Dans notre application, les points d'intérêt sont majoritairement définis par des zones de 16×16 pixels.

Ces constatations montrent que la correction basée sur un obturateur interne est aussi efficace que la méthode basée sur l'étalonnage si l'on considère le bruit sur des fenêtres dont la taille est de l'ordre de 20× 20 pixels. L'avantage de l'obturateur est qu'il ne nécessite pas une longue période d'étalonnage. Cependant, comme cela est évoqué dans les références [84,86], lorsque l'obturateur mécanique est déclenché pour mettre à jour l'offset, la caméra devient non-opérationnelle. De plus, l'obturateur mécanique augmente le prix et le poids de la caméra.

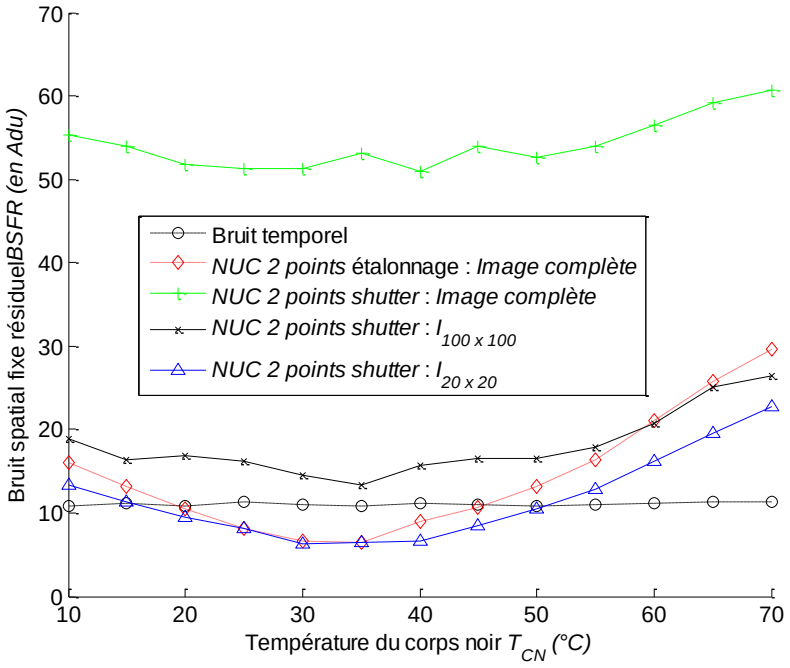


Figure 58. Bruit spatial fixe résiduel *BSFR* en fonction de la température du corps noir T_{CN}. Deux corrections *NUC 2 points* sont comparées sur ce graphique. La première est obtenue par étalonnage (courbe en pointillés rouge avec des losanges). La seconde est obtenue grâce au *shutter* et le *BSFR* est calculé sur l'image complète (courbe en trait plein noir avec des +), sur une fenêtre de 100×100 pixels (courbe en trait plein noir avec des ×) et sur une fenêtre de 20×20 pixels (courbe en trait plein bleu avec des triangles).

On trouve dans la littérature des moyens de se référer à un corps uniforme qui ne bloque pas le rayonnement provenant de la scène. Ainsi le système n'est jamais in-opérationnel même durant les phases de mise à jour de l'offset. Cependant ces techniques nécessitent d'introduire du matériel couteux dans le système global. Dans la référence [88] les auteurs proposent d'utiliser une source de température modulée en fréquence et éloignée de la distance de mise au point du système optique. Dans le même esprit, les mêmes auteurs proposent dans la référence [89] d'introduire un miroir semi-transparent en chalcogénure (ou l'un de ses dérivés) orienté à 45° avec l'axe optique. Un corps de référence peut ainsi être introduit à la perpendiculaire de l'axe optique tout en étant visible grâce au miroir à 45°.

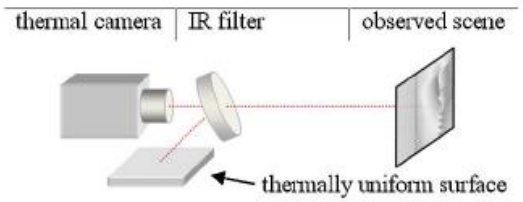


Figure 57. Utilisation d'un miroir semi-réfléchissant pour se référer à un corps uniforme permettant de rafraichir la *NUC* sans bloquer le rayonnement de la scène [89].

D'autres méthodes basées sur des algorithmes de traitement d'images ont également été développées et ne nécessitent ni obturateur, ni longue phase d'étalonnage. Nous allons les présenter et les tester dans la section suivante.

2.5.6. Les corrections *shutterless* basées sur la scène *(scene-based)*

2.5.6.1. Méthodes basées sur les statistiques temporelles constantes

La réponse d'un pixel est considérée comme une variable aléatoire $y(t)$ qui dépend du temps t. Elle s'exprime en fonction de la réponse idéale du détecteur $z(t)$, d'un gain a et d'un offset b. Un fois de plus nous cherchons à estimer $z(t)$.

$$y(t) = az(t) + b \tag{2.17}$$

Nous ferons remarquer au lecteur que ces notations ne sont pas en gras car, pour simplifier l'explication, nous nous concentrons sur un seul pixel.

Dans la référence [90] Harris & *al.* définissent la contrainte des statistiques constantes (dans la littérature anglophone on parle d'algorithmes *constant-statisitcs cs*). Cette contrainte utilise deux hypothèses :

- les moyennes temporelles m_z des réponses idéales des pixels sont identiques pour tous les pixels de la matrice,
- les variances temporelles σ_z^2 des réponses idéales des pixels sont identiques pour tous les pixels de la matrice.

Cette hypothèse implique que la caméra est en mouvement relatif par rapport à la scène et que l'on considère la moyenne temporelle m_z et la variance temporelle σ_z^2 des pixels sur une longue période.

Pour estimer l'offset b, exprimons la réponse moyenne temporelle d'un pixel :

$$m_y = E[az(t) + b] = am_z + b \tag{2.18}$$

En considérant que $m_z = 0$ (cette considération n'impacte pas la correction finale à un facteur d'échelle globale près, au regard de la non-uniformité de l'image), on obtient directement l'offset :

$$\hat{b} = m_y \tag{2.19}$$

Pour estimer le gain a, exprimons la variance temporelle d'un pixel :

$$\sigma_y^2 = var[az(t) + b] = a\sigma_z^2 \tag{2.20}$$

Comme nous considérons que la variance temporelle σ_z^2 est identique pour tous les pixels de la matrice, à un facteur d'échelle près (ceci n'impacte pas la correction finale au regard de la non-uniformité de l'image), alors le gain peut être estimé comme suit :

$$\hat{a} = \sigma_y^2 \tag{2.21}$$

Finalement grâce aux hypothèses faites dans le cadre de la contrainte des statistiques constantes *cs* la correction *2 points* de la non-uniformité *NU* s'exprime comme suit :

$$\hat{z} = \frac{y - \hat{b}}{\hat{a}} \tag{2.22}$$

Dans la référence [91] la contrainte des statistiques constantes *cs* est également utilisée avec des caméras fonctionnant dans la gamme *MWIR* (3-5 µm). Une première étape permet de traiter la non-uniformité *NU* qui provient de l'architecture du circuit de lecture. Il est considéré que des groupes de pixels possèdent un offset et un gain commun car souvent, les colonnes (ou les lignes en fonction de l'architecture électronique) d'une matrice de pixels partagent un même amplificateur. La contrainte *cs* est utilisée pour imposer une moyenne spatio-temporelle et une variance spatio-temporelle identiques à tous les groupes (un groupe peut être une ligne ou une colonne) de pixels. Cette opération utilise l'accumulation temporelle des images pour converger vers une solution satisfaisante grâce à un filtre récursif. Après cette étape, on note la valeur d'un pixel $\hat{y}$.

Puis, toujours dans la référence [91], une seconde étape permet de corriger la non-uniformité *NU* à l'échelle du pixel en utilisant un filtre médian de taille 5×5 pixels. La réponse à ce filtre pixels est noté $\bar{z}$. En utilisant $n = 100$ images au cours desquelles il y a eu un mouvement relatif entre la caméra et la scène, les auteurs déterminent le gain et l'offset des pixels qui minimisent la différence du carré entre la valeur du pixel après le filtrage médian $\bar{z}$ et la valeur du pixel après la correction de la partie du *BSF* due à l'architecture électronique (après cette correction due à l'architecture, le pixel est noté $\hat{y}$).

$$\arg\min_{a} \sum_{i=1}^{n} \lambda^{n-i+1} |\hat{y}_i - \boldsymbol{a}^T \bar{\boldsymbol{z}}_i|^2 \tag{2.23}$$

Le nombre d'images accumulées est noté n. La valeur du pixel après correction du bruit du à l'architecture du circuit de lecture à la $i^{ème}$ image est noté $\hat{z}_i$. La valeur du pixel après filtrage médian à la $i^{ème}$ image est noté $\bar{z}_i$. Le vecteur $\bar{\boldsymbol{z}}_i$ correspond à $\bar{\boldsymbol{z}}_i = [\bar{z}_i, 1]^T$ et le vecteur $\boldsymbol{a}$ est composé du gain et de l'offset de correction que nous recherchons $[a, b]^T$. Le facteur λ^{n-i} lorsque $\lambda < 1$ permet de donner plus de poids aux images les plus récentes pour éviter les problèmes d'effet de *ghost*.

2.5.6.2. Recalage d'images

Le recalage d'images (*image registration* dans la littérature anglophone) consiste à estimer la transformation géométrique qui permet de passer d'une image à une autre, c'est-à-dire le changement de point de vue. On trouve dans la littérature des techniques utilisant le fait que le changement de point de vue entre plusieurs images successives est connu (ou peut être estimé à partir des images brutes). Dans la référence [92] les auteurs ont travaillé avec des caméras sensibles au rayonnement *MWIR* (3-5 µm). Ils considèrent que le recalage d'images est possible lorsque les non-uniformités sont relativement faibles. Pour estimer ce mouvement, ils considèrent que l'objet est loin de la caméra et que le mouvement est faible d'une image à l'autre. Ainsi, seulement trois paramètres sont estimés et permettent de décrire le mouvement

de la caméra : la rotation dans le plan image et deux translations dans le plan image. Finalement, au cours d'une séquence d'images, plusieurs pixels *'voient'* la même zone de la scène. Les auteurs font l'hypothèse que la moyenne de ces pixels est une bonne estimation de la réponse corrigée de la non-uniformité *NU*.

2.5.6.3. Méthodes fonctionnant à partir d'une seule image

Dans la référence [93] une méthode de correction du *BSF* basée sur des moyennes non locales *NL-means algorithm*, est proposée. Considérons la valeur d'un pixel de l'image $y(i)\ \forall\ i\ \epsilon\ I$. L'estimation de la valeur corrigée $\hat{y}(j)$ est la moyenne pondérée de tous les pixels de l'image :

$$\hat{y}(i) = \sum_{j\epsilon I} w(i,j)y(j) \tag{2.24}$$

L'ensemble des poids $w(i,j)$ est défini tel que $0 < w(i,j) < 1$ et $\sum_{j\epsilon I} w(i,j) = 1$. Ils sont calculés à partir d'un critère de similarité basé sur la distance euclidienne entre les patchs carrés centrés sur i et j. Comme cela est évoqué dans la référence [94], les méthodes de correction du *BSF* basées sur le calcul de moyennes lissent l'image et, même si le résultat semble intéressant à l'œil, les informations hautes fréquences sont perdues. L'algorithme *NL-means* ne déroge pas à cette règle.

D'autres travaux proposent de corriger le bruit de colonnes [94,95,96,97,98]. Dans la référence [94] les auteurs modélisent le bruit de colonne comme un bruit aditif :

$$Y(i,j) = Z(i,j) + S_j(i) \tag{2.25}$$

La valeur d'un pixel de l'image brute à la $i^{ème}$ ligne et à la $j^{ème}$ colonne est $Y(i,j)$**.** La réponse idéale de ce même pixel est $Z(i,j)$. Tous les pixels appartenant à la $j^{ème}$ colonne sont entachés d'un bruit $S_j(i)$ que les auteurs jugent pertinent (d'après une expérience décrite dans l'article) de modéliser par un polynôme qui dépend de la réponse brute du pixel **:**

$$S_j(i) = a_k\left(\left(Y(i,j)\right)_j\right)^2 + b_k\left(Y(i,j)\right)_j + c_k \tag{2.26}$$

La notation $\left(Y(i,j)\right)_j$ signifie que la variable j est fixée. Ils partent ensuite de l'*a priori* que l'image de la scène est lisse et que le bruit de colonne possède une haute fréquence spatiale. Donc un filtrage passe bas 1D w permet d'extraire, majoritairement le bruit de colonne :

$$Y_D(i,j) = Y(i,j) - \sum_{\tau} w\,(j+\tau)Y(i,j) \tag{2.27}$$

L'image $\boldsymbol{Y}_D$ contient les détails haute fréquence de l'image. Ensuite les coefficients du polynôme qui modélisent le bruit de colonne sont recherchés en minimisant l'écart quadratique entre l'image brute $\boldsymbol{Y}$ et l'image des détails hautes fréquences $\boldsymbol{Y}_D$:

$$\arg\min_{a_k,b_k,c_k} \sum_i \left(a_k\left(\left(Y(i,j)\right)_j\right)^2 + b_k\left(Y(i,j)\right)_j + c_k - \left(Y_D(i,j)\right)_j\right)^2 \tag{2.28}$$

Le principe de l'algorithme est récapitulé sur la Figure 59.

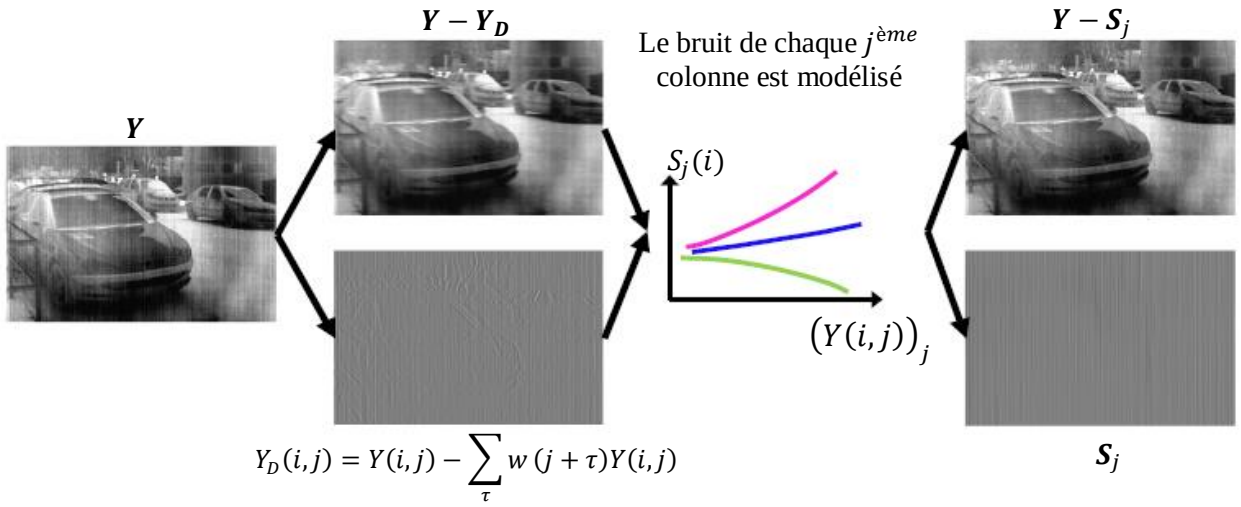


Figure 59. Principe de la correction du bruit de colonne de la référence [94].

La référence [95] propose une autre méthode, appelée *MIRE* (*midway infrared equalization*), pour corriger le bruit de colonne. Celle-ci exploite l'idée que la scène est continue et qu'elle possède peu de bords. Par conséquent la différence entre deux colonnes adjacentes est statistiquement petite. Les auteurs proposent d'égaliser les histogrammes cumulatifs des colonnes adjacentes. Parmi les différentes méthodes qui permettent d'égaliser des histogrammes, celle appelée en littérature anglophone *midway equalization* est utilisée [99]. Si l'on considère deux histogrammes cumulatifs H_1 et H_2 correspondant à deux colonnes adjacentes, l'histogramme égalisé H_{mid} s'exprime comme suit :

$$H_{Mid}^{-1} = \frac{H_1^{-1} + H_2^{-1}}{2} \tag{2.29}$$

En cas de fort bruit, plus de deux colonnes voisines sont utilisées pour calculer la valeur de l'histogramme. Dans ce cas, un poids gaussien est attribué aux histogrammes des colonnes voisines :

$$H_j^{-1} = \sum_{k\epsilon(-n,\dots,+n)} g(k) H_{k+j}^{-1} \tag{2.30}$$

La colonne à corriger est la colonne j. Les colonnes voisines sont notées k avec $k\epsilon(-n, \dots, +n)$. La fonction de poids gaussien s'exprime ainsi :

$$g(k) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{k^2}{2s^2}} \tag{2.31}$$

L'écart type s de cette fonction gaussienne est un paramètre que les auteurs proposent de déterminer automatiquement. Soit $\boldsymbol{I}$ l'image corrigée grâce à cette méthode, et $I_{i,j}$ la valeur du pixel à la $i^{ième}$ ligne et à la $j^{ème}$ colonne. La fonction de variation des lignes *TV-line* (*total variation-line*) de l'image est à minimiser :

$$\|I\|_{TV\_line} = \sum_{i,j} |I_{i,j+1} - I_{i,j}|. \tag{2.32}$$

Le paramètre s qui minimise la fonction $\|I\|_{TV\_line}$ permet d'obtenir de bons résultats.

$$s = \arg\min_{s}\left(\|I\|_{TV\_line}\right), \tag{2.33}$$

Nous avons testé la méthode *MIRE* sur des images acquises à différentes températures de la caméra T_C (Figure 60). Chaque colonne de la première ligne de la Figure 60 représente une image brute acquise à une température de caméra donnée. De la gauche vers la droite, la température de la caméra vaut 21.5°C, 38.03°C et 40.84°C. Sachant qu'un cube d'images $\boldsymbol{C}(T_{CN}, T_C)$ (ou une simple image $\boldsymbol{O}(T_{CN}, T_C)$) d'un corps noir devra probablement être acquis lors de la fabrication d'une caméra thermique pour contrôler la qualité du produit, nous avons trouvé pertinent d'exploiter ces données. Ainsi l'image brute a été dans un premier temps corrigée grâce à la moyenne temporelle du cube d'images d'un corps noir $\overline{\boldsymbol{C}}(T_{CN}, T_C)$ acquis à une température de caméra $T_C = 47.14°C$ (la température de l'ambiante était alors de 35°C) et face à un corps noir à la température $T_{CN} = 30°C$:

$$\boldsymbol{Y}_1 = \boldsymbol{Y} - \overline{\boldsymbol{C}}(T_{CN} = 30°C, T_C = 47.14°C) \tag{2.34}$$

Le résultat de cette première correction est illustré sur la seconde ligne de Figure 60. Très logiquement, les images acquises à des températures (de la caméra) éloignées de la température de la caméra lors de l'acquisition du cube d'images du corps noirs $\overline{\boldsymbol{C}}(T_{CN} = 30°C, T_C = 47.14°C)$ sont les plus bruitées.

On applique ensuite à $\boldsymbol{Y}_1$ l'algorithme *MIRE* pour réduire le bruit de colonne et nous obtenons les images $\boldsymbol{Y}_2$ représentées à la troisième ligne de la Figure 60. On constate que le bruit de colonne est réduit. Cependant un bruit à l'échelle du pixel demeure ainsi qu'un bruit basse fréquence.

L'algorithme *NL-means*, basé sur un lissage par *patch*, est capable de corriger le bruit à l'échelle du pixel. Cependant, comme nous l'avons dit plus haut, des informations hautes fréquences vont être perdues. Nous appliquons quand même l'algorithme *NL-means* et nous obtenons la quatrième ligne de la Figure 60. Nous constatons qu'un bruit basse fréquence persiste et semble très gênant (particulièrement sur l'image de gauche de la quatrième ligne de la Figure 60).

D'autres travaux exploitant encore plus le cube d'images qui pourrait être acquis lors d'une phase de contrôle à l'issue de la fabrication de la caméra ont été menés et sont présentés ci-après. Ces travaux améliorent la correction du bruit basse fréquence résiduel.

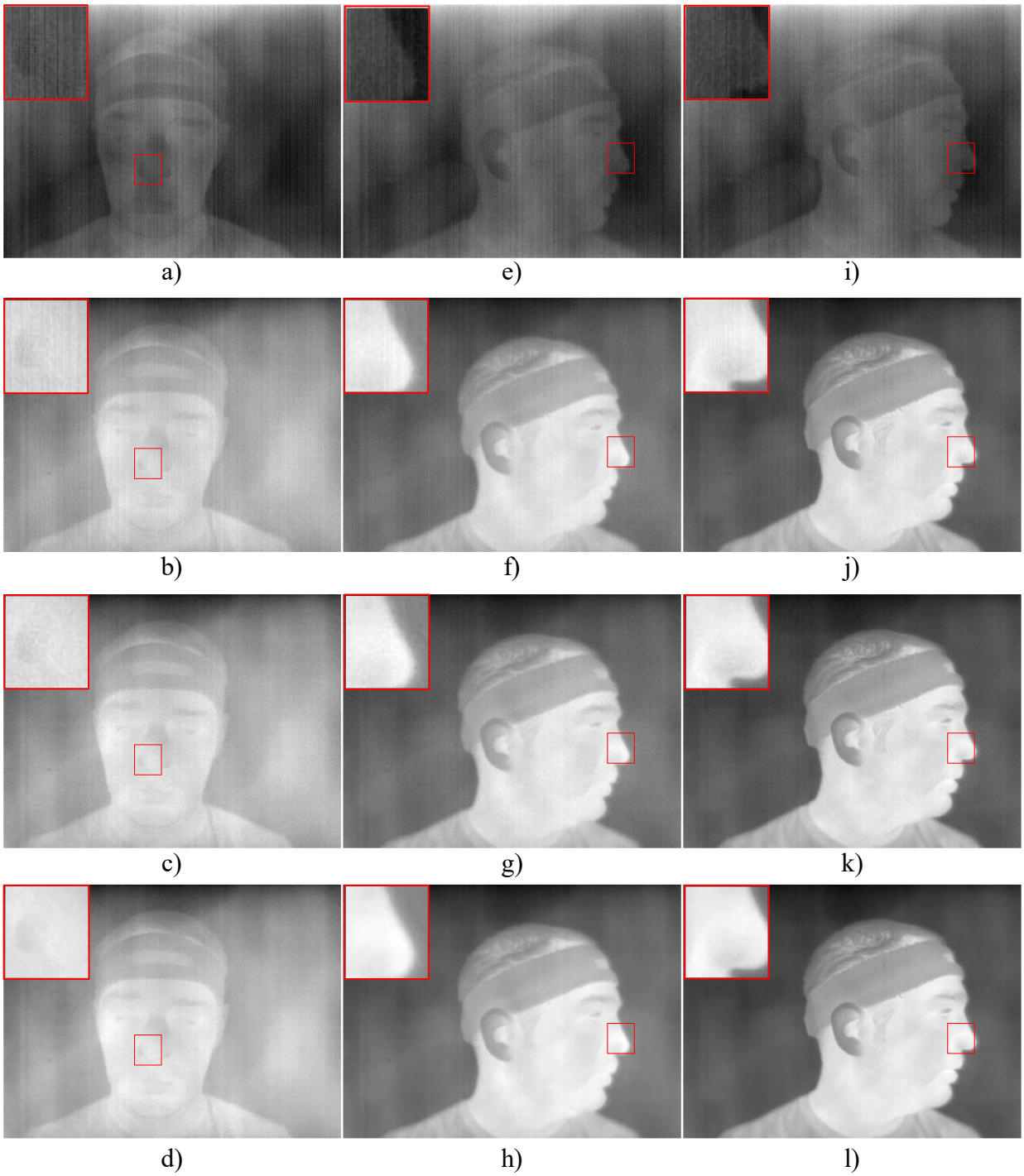


Figure 60. La température de la camera T_C varie suivant les colonnes. De a) à d) la température de la caméra est $T_C = 21.50°C$. De e) à h) la température de la caméra T_C vaut $38.03°C$. De i) à l) a température de la caméra est T_C vaut $40.84°C$. Sur la première ligne, les images $\boldsymbol{Y}$ sont brutes. Sur la seconde ligne, les images $\boldsymbol{Y}_1$ sont le résultat d'une soustraction de l'image brute avec la moyenne temporelle d'un cube d'images du corps noir à $T_{CN} = 30°C$ acquis pour une température de caméra à $T_C = 47.14°C$. Sur la troisième ligne, on applique aux images $\boldsymbol{Y}_1$ l'algorithme *MIRE* et on obtient les images $\boldsymbol{Y}_2$. Enfin, sur la dernière ligne, on applique aux images $\boldsymbol{Y}_2$ l'algorithme *NL-means* et on obtient les images $\boldsymbol{Y}_3$.

La référence [100] est une interprétation du brevet détenu par la société *Ulis*. Nous ne reviendrons pas sur l'algorithme complet mais nous expliquerons les premières étapes pour en comprendre l'esprit. Nous les implémenterons et les testerons sur des images réelles. Les auteurs proposent de distinguer le bruit de colonne $\boldsymbol{B}_{colonne}$ du reste du bruit appelé bruit de dispersion $\boldsymbol{B}_{disp}$. Leur méthode est basée sur l'utilisation d'au moins une image du corps noir acquise lors de la phase de contrôle de la caméra alors que la caméra est à une température T_C donnée. Rappelons que l'image d'un corps noir correspond à la carte du *BSF* du détecteur. Ainsi, cette carte de bruit notée $\boldsymbol{B}$ (au format 640 colonnes × 480 lignes) correspond à la somme du bruit de colonne $\boldsymbol{B}_{colonne}$ (au format 640 colonnes × 480 lignes) et du bruit de dispersion $\boldsymbol{B}_{disp}$ (au format 640 colonnes ×480 lignes) :

$$\boldsymbol{B} = \boldsymbol{B}_{disp} + \boldsymbol{B}_{colonne} \tag{2.35}$$

Pour extraire la composante de bruit de colonne, la moyenne de chaque colonne de l'image du corps noir $\boldsymbol{B}$ est calculée, ce qui donne un vecteur $\boldsymbol{V}_{colonne}$ (au format 640 colonnes × 1 ligne). La matrice $\boldsymbol{B}_{colonne}$ est donc la répétition du vecteur $\boldsymbol{V}_{colonne}$ le long des 480 lignes. Il est ensuite trivial de déduire la matrice du bruit de dispersion $\boldsymbol{B}_{disp} = \boldsymbol{B} - \boldsymbol{B}_{colonne}$. Les matrices $\boldsymbol{B}_{colonne}$ et $\boldsymbol{B}_{disp}$ viennent d'être définies de manière rigoureuse pour une température T_C. Les auteurs font l'hypothèse que ces contributions sont identiques à des températures T_C différentes, à un facteur multiplicatif global près. La suite de leur méthode consiste à rechercher ces facteurs multiplicatifs.

L'image brute à corriger est toujours notée $\boldsymbol{Y}$ et nous noterons $\boldsymbol{V}_{\boldsymbol{Y}}$ le vecteur dont les éléments correspondent à la moyenne des colonnes de $\boldsymbol{Y}$ ($\boldsymbol{V}_{\boldsymbol{Y}}$ a un format de 640 colonnes × 1 ligne). Les auteurs proposent une formulation type problème inverse. L'objectif est de décrire le bruit de colonne de l'image brute $\boldsymbol{V}_{\boldsymbol{Y}}$ par le bruit de colonne extrait de l'image du corps noir $\boldsymbol{B}_{colonne}$. Ils proposent ainsi d'estimer un paramètre multiplicatif α grâce à la méthode des moindres carrés :

$$\arg\min_{\alpha \in \mathbb{R}} \sum\nolimits_{640\ colonnes} (\boldsymbol{V}_{\boldsymbol{Y}} - \alpha \boldsymbol{V}_{colonne})^2 \tag{2.36}$$

Les auteurs proposent d'appliquer un filtre passe haut 1D à $\boldsymbol{V}_{\boldsymbol{Y}}$ et à $\boldsymbol{V}_{colonne}$ avant d'appliquer la méthode des moindres carré car ils font l'hypothèse que la scène est continue et possède statistiquement peu de hautes fréquences. La solution optimale à ce type de problème inverse est :

$$\alpha = \frac{\sum_{640\ colonnes}(\boldsymbol{V}_{\boldsymbol{Y}} \times \boldsymbol{V}_{colonne})}{\sum_{640\ colonnes}(\boldsymbol{V}_{colonne}^2)} \tag{2.37}$$

Lorsque le paramètre α est déterminé, on applique la correction suivante à l'image brute :

$$\boldsymbol{Y}_1 = \boldsymbol{Y} - \alpha \boldsymbol{B}_{colonne} \tag{2.38}$$

L'image $\boldsymbol{Y}_1$ est corrigée du bruit de colonne.

Toujours dans la référence [100] après avoir corrigé le bruit de colonne, les auteurs proposent une seconde étape pour corriger le bruit de dispersion. Là encore, une formulation type problème inverse vise à décrire, grâce à un facteur multiplicatif β, le bruit de dispersion par $\beta \times \boldsymbol{B}_{disp}$. Pour cela le paramètre β est estimé en minimisant la variation totale de l'image corrigée. La variation totale *TV* de l'image corrigée étant :

$$TV(\boldsymbol{Y}_1 - \beta \boldsymbol{B}_{disp}) = \sum_{640\ colonnes,480\ lignes} \left|\boldsymbol{\nabla}.(\boldsymbol{Y}_1 - \beta \boldsymbol{B}_{disp})\right|^2 \quad (2.39)$$

avec l'opérateur gradient défini comme suit :

$$\boldsymbol{\nabla} = \begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{bmatrix}$$

Le paramètre β s'exprime finalement comme suit :

$$\beta = \arg\min_{\beta \epsilon \mathbb{R}} \sum_{640\ colonnes,480\ lignes} \left|\boldsymbol{\nabla}.(\boldsymbol{Y}_1 - \beta \boldsymbol{B}_{disp})\right|^2 \quad (2.40)$$

Ce problème possède une solution optimale qui est :

$$\beta = \frac{\sum_{640\ colonnes,480\ lignes}(\nabla_{colonne}.\boldsymbol{Y}_1.\nabla_{colonne}.\boldsymbol{B}_{disp} + \nabla_{ligne}.\boldsymbol{Y}_1.\nabla_{ligne}.\boldsymbol{B}_{disp})}{\sum_{640\ colonnes,480\ lignes}\left((\nabla_{colonne}.\boldsymbol{B}_{disp})^2 + (\nabla_{ligne}.\boldsymbol{B}_{disp})^2\right)} \quad (2.41)$$

Lorsque le paramètre β est déterminé, on applique la correction suivante à l'image $\boldsymbol{Y}_1$ déjà corrigée du bruit de colonne :

$$\boldsymbol{Y}_2 = \boldsymbol{Y}_1 - \alpha \boldsymbol{B}_{disp} \quad (2.42)$$

On obtient l'image corrigée finale $\boldsymbol{Y}_2$.

Nous avons implémenté ces premières étapes du brevet sur des images acquises à différentes températures de la caméra T_C (cf. Figure 61). Chaque colonne de la première ligne de la Figure 61 représente une image brute $\boldsymbol{Y}$ acquise à une température de caméra donnée. De la gauche vers la droite, la température de la caméra vaut 21.5°C, 38.03°C et 40.84°C. Nous avons appliqué une interprétation de la correction détaillée dans le brevet [100] aux images brutes $\boldsymbol{Y}$. L'image de référence du corps noir permettant de définir $\boldsymbol{B}_{colonne}$ et $\boldsymbol{B}_{dispersion}$ est la moyenne temporelle du cube d'images du corps noir : $\overline{\boldsymbol{C}}(T_{CN} = 30°C, T_C = 47.14°C)$. Les images $\boldsymbol{Y}_1$ obtenues après la correction du bruit de colonne $\boldsymbol{B}_{colonne}$ sont représentées sur la seconde ligne de la Figure 61. Nous avons filtré $\boldsymbol{V}_{colonne}$ et $\boldsymbol{V}_{\boldsymbol{Y}}$ avec le filtre passe haut $\boldsymbol{F}$ avant de calculer le paramètre α grâce à l'équation (2.37) comme cela est évoqué dans le brevet [100].

$$\boldsymbol{F} = [0.0456 \quad -0.0288 \quad -0.2956 \quad 0.5575 \quad 0.0456 \quad -0.0288 \quad -0.2956]$$

Les images $\boldsymbol{Y}_2$ obtenues après correction du bruit de dispersion $\boldsymbol{B}_{dispersion}$ sont représentées sur la troisième ligne de la Figure 61.

On constate qu'après avoir appliqué la correction issue de notre interprétation de ces premières étapes, un bruit spatial résiduel de colonne important demeure. Nous avons appliqué l'algorithme *MIRE* à l'image $\boldsymbol{Y}_2$ pour réduire ce bruit de colonne résiduel. On obtient l'image $\boldsymbol{Y}_3$ représentée sur la quatrième ligne de la Figure 61.

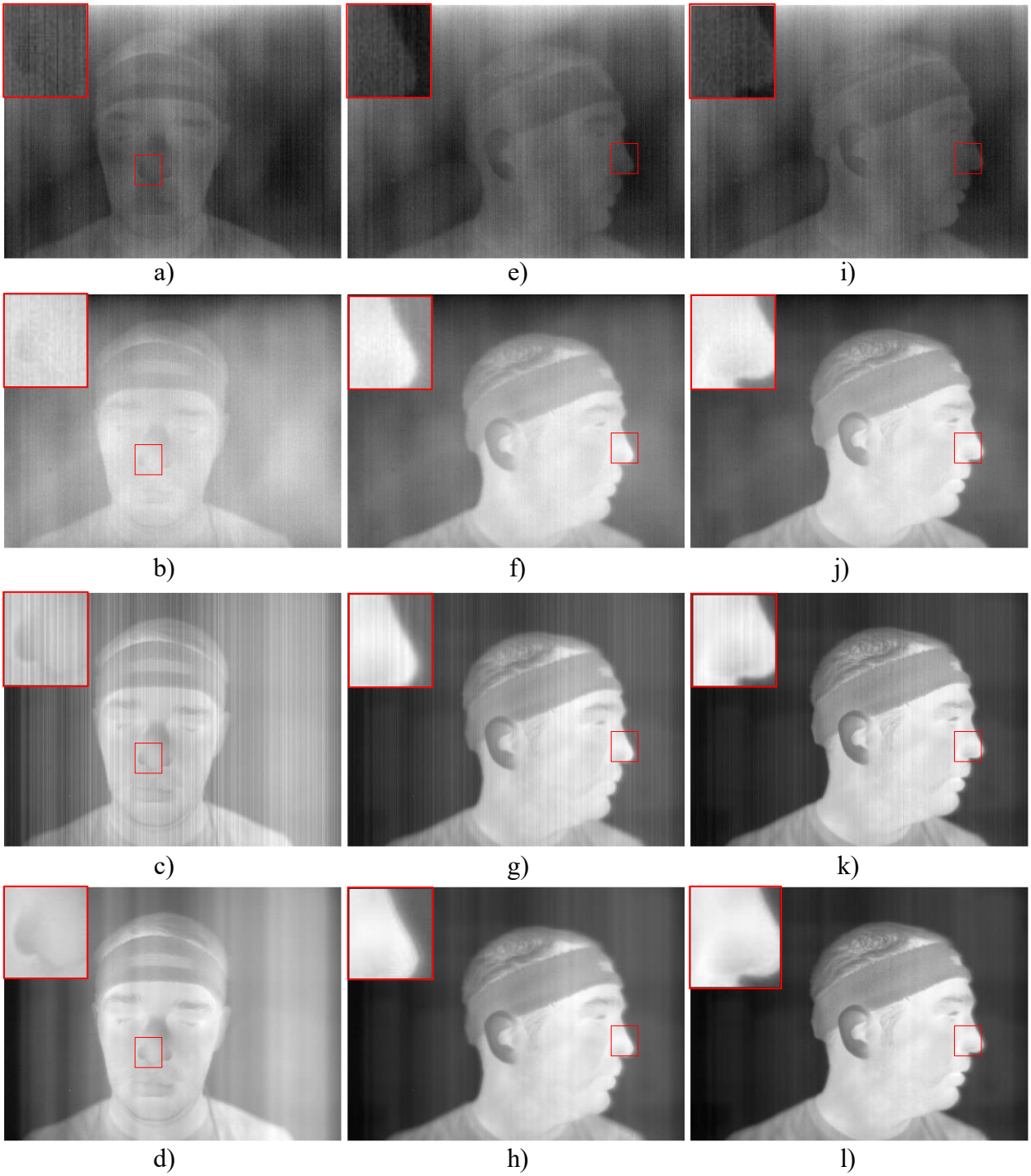


Figure 61. La température de la camera T_C varie suivant les colonnes. De a) à d) la température de la caméra est $T_C = 21.50°C$. De e) à h) la température de la caméra T_C vaut $38.03°C$. De i) à l) a température de la caméra T_C vaut $40.84°C$. Sur la première ligne, les images $\boldsymbol{Y}$ sont brutes. Sur la seconde ligne, on a simplement soustrait la moyenne d'un cube d'images du corps noir acquis à $T_{CN} = 30°C$ et $T_C = 47.14°C$. Sur la troisième ligne, les images $\boldsymbol{Y}_2$ (cf. équation 2.1.1.1(2.42)) ont été obtenues avec notre implémentation des premières étapes du brevet [100] qui utilise un cube d'images du corps noir à $T_{CN} = 30°C$ acquis pour une température de caméra $T_C = 47.14°C$. Enfin, sur la dernière ligne, on applique aux images $\boldsymbol{Y}_2$ l'algorithme *MIRE* et on obtient les images $\boldsymbol{Y}_3$.

Comme nous n'avons pas implémenté l'algorithme de manière complète, nous avons demandé à la société *Ulis* de corriger quelques images pour avoir un aperçu plus représentatif des capacités de cette méthode. Une version améliorée, mais néanmoins ni optimal, ni complète a été testée avec nos images. Nous leur avons fourni l'image brute acquise à $T_C = 21.50°C$ et deux cubes d'images du corps noir à 30°C et à 40°C pour une température de caméra $T_C = 47.14°C$:

$\boldsymbol{C}(T_{CN} = 30°C, T_C = 47.14°C)$, et $\boldsymbol{C}(T_{CN} = 40°C, T_C = 47.14°C)$.

Le résultat obtenu est comparé à celui obtenu avec notre implémentation (cf. Figure 62). Les résultats obtenus avec l'implémentation, non complète et non optimale, de la société Ulis donne une image qui visuellement parait très convenable.

a) b)

Figure 62. Comparaison de notre implémentation non-optimisée du brevet [100] et de celle, plus complète, réalisée par les auteurs. a) Notre implémentation basique du brevet utilisant un cube d'images du corps noir à $T_{CN} = 30°C$ à la température caméra $T_C = 47.14°C$ et l'algorithme *MIRE* pour supprimer les colonnes résiduelles. b) Une autre implémentation plus complexe (mais toujours pas complète) du brevet [100] utilisant deux cubes d'images du corps noir, l'un à $T_{CN} = 30°C$, l'autre à $T_{CN} = 40°C$, tous deux acquis pour une température caméra $T_C = 47.14°C$.

Dans l'implémentation partielle qu'il nous a été autorisée de tester, l'algorithme prend en compte un gain de correction qui peut être adapté via un facteur multiplicatif recherché automatiquement par une méthode de minimisation similaire à celles qui ont permis d'estimer les coefficients α et β.

Un traitement supplémentaire permet également de traiter les artefacts de colonne. Pour cette étape, le calcul d'un poids basé sur la probabilité qu'un pixel de l'image soit dans une zone uniforme est nécessaire. Ensuite la différence pondérée entre un pixel et ceux des colonnes voisines permet d'établir la valeur de correction résiduelle de colonne à appliquer.

Remarque : comme nous le voyons sur la troisième ligne de la Figure 61, lorsque nous appliquons notre implémentation des premières étapes du brevet (cf. équations (2.38) et (2.42)), nous obtenons un bruit de colonne résiduel très important. Nous avons appliqué la correction MIRE spécialement conçu pour

réduire le bruit de colonne (on obtient la quatrième ligne la Figure 61). Le principe de correction du bruit de colonne résiduel du brevet Ulis semble plus abouti, mais nous ne l'avons pas implémenté.

2.5.7. Conclusion sur les différentes méthodes de correction du bruit spatial

Nous l'avons vu, la *NUC* doit être rafraichie régulièrement lorsque la température T_C évolue. Ceci est dû essentiellement à la dépendance de l'offset $\boldsymbol{B}$ à T_C. Cette dépendance peut être gérée de trois manières différentes :

- par étalonnage dans une chambre climatique suivi d'une régression polynomiale en fonction de la température de la caméra T_C (approche coûteuse en temps mais très performante),
- grâce à un obturateur mécanique (approche classique qui nécessite une pièce mécanique motorisée, et que l'image soit figée pendant la mise à jour de la *NUC*),
- en utilisant des algorithmes de traitement d'images tels que des filtres spatio-temporels combinés à une table d'étalonnage usine judicieusement exploitée comme dans l'interprétation du brevet [100] (approche bas coût).

La Figure 63 illustre quatre méthodes de correction (expliquées dans les sections précédentes) d'une image brute acquise à 21.50°C. Les deux images de la première ligne de la Figure 63 illustrent des méthodes de correction basées sur des algorithmes de traitement d'images et une image de référence d'un corps noir à 30°C acquis lorsque la caméra est à 47.14°C. L'image a) est obtenue en utilisant la soustraction de l'image de référence, l'algorithme *MIRE* puis l'algorithme *NL-means*. L'image b) est obtenue grâce à une interprétation du brevet [100] par ses auteurs. La seconde ligne de la Figure 63 illustre la correction basée sur l'obturateur mécanique (image c) et sur l'étalonnage (image d).

La solution basée sur l'obturateur mécanique est avantageuse car elle ne nécessite pas d'étalonnage, et de plus elle est efficace à la vue de l'estimation du bruit spatial fixe résiduel *BSFR* sur des fenêtres de pixels au format 20×20 (cf. Figure 58) quel que soit la température de la caméra. L'inconvénient majeur est que le système global est non-opérationnel lorsque l'obturateur mécanique est déclenché pour mettre à jour l'offset. De plus comme souvent évoqué dans la littérature scientifique, l'obturateur mécanique peut casser et augmente le poids du système [84,86].

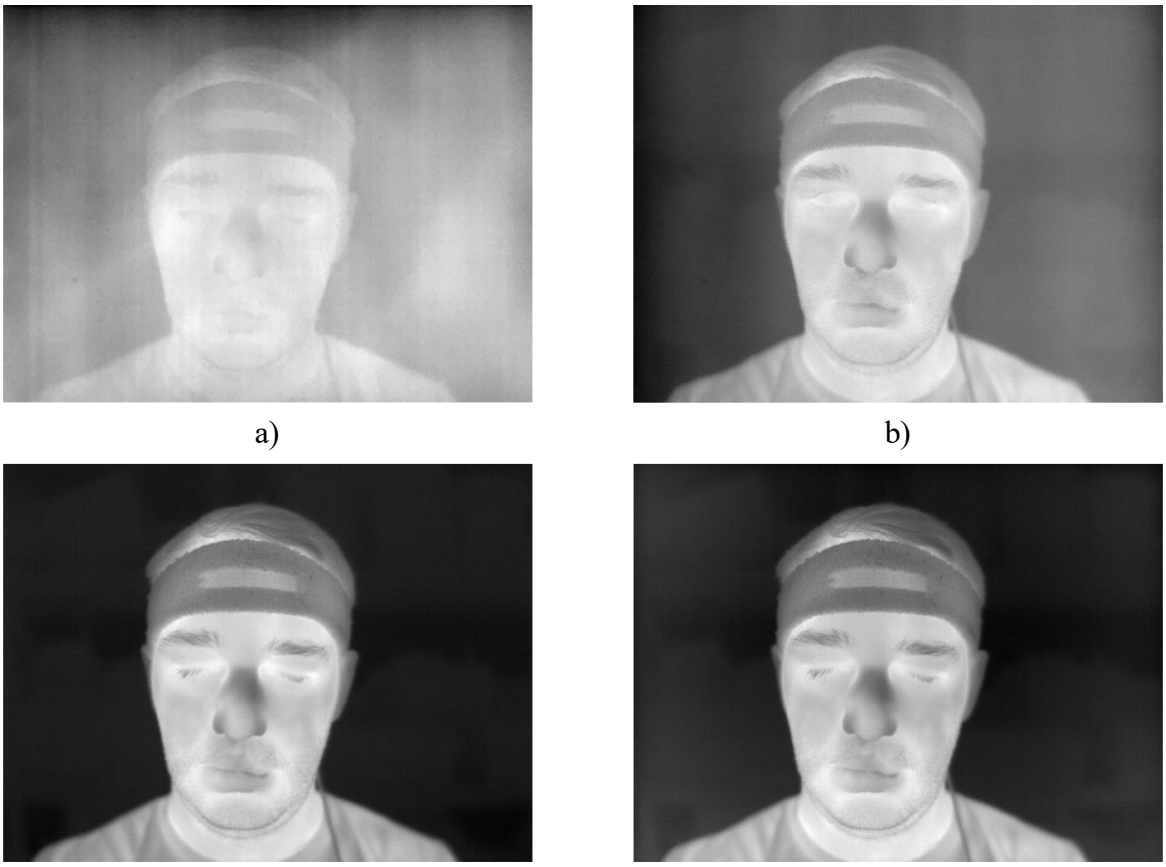
L'étalonnage est la méthode la plus efficace mais c'est également une méthode extrêmement contraignante à mettre en œuvre car dans notre implémentation, nous avons réalisé un étalonnage de 13h en chambre climatique.

Les méthodes de correction du bruit basées sur des algorithmes de traitement d'images et quelques images de référence sont avantageuses car elles ne nécessitent ni obturateur mécanique, ni étalonnage de longue durée. Elles permettent d'obtenir des images agréables à regarder à l'œil lorsque la température de la caméra n'évolue pas trop par rapport à la température lors de l'acquisition de l'image (ou du cube d'images) de référence (comme sur la Figure 63 où cette différence de température vaut environ 17°C). Si l'on considère les progrès conjoints des matrices de microbolomètres et des méthodes de traitements

d'images, il semble raisonnable d'imaginer des systèmes basés sur une matrice de microbolomètres *shutterless* ne nécessitant pas d'étalonnage long.

Nous avons estimé qu'il était plus judicieux de développer notre application d'estimation de la pose 3D du visage à partir des images les mieux corrigées : c'est-à-dire les images obtenues grâce aux corrections type obturateur mécanique ou étalonnage. Tout en sachant que ces méthodes ne sont peut-être pas compatibles avec l'intégralité des contraintes automobiles, et que les algorithmes *shutterless* progressent.

Nous avons présenté, détaillé et testé certaines des méthodes de l'état de l'art permettant de corriger la non-uniformité *NU* de l'image. L'évolution de la température de la caméra T_C, au-delà de la modification de la non-uniformité *NU* de l'image, engendre une évolution des réponses des pixels. Nous allons mettre ceci en parallèle avec les algorithmes d'estimations de la pose 3D du visage et montrer que cela peut poser problème. Puis nous discuterons d'une méthode possible pour compenser la dépendance à la température T_C de la caméra du niveau moyen de l'image grâce à des pixels de référence.



a) b)

c) d)

Figure 63. Illustration de quatre méthodes de correction du bruit spatial. La température de la caméra lors de l'acquisition de l'image brute est $T_C = 21.5°C$. L'image a) a été obtenue grâce à l'algorithmes *MIRE* suivi de l'algorithme *NL-means*. L'image b) a été obtenue grâce au brevet de la société Ulis. L'image c) a été obtenue grâce une correction *2 points* basée sur un gain usine et un offset enregistré grâce à un obturateur mécanique interne. L'image d) a été obtenue grâce à une correction *2 points* obtenue par un étalonnage de 13 heures en chambre climatique.

2.6. Etalonnage radiométrique

2.6.1. Introduction

Nous venons de discuter dans les sections précédentes de différentes méthodes pour corriger les non-uniformités *NUs* de l'image. Ces méthodes ne permettent pas de donner un sens radiométrique aux réponses des pixels. Pour nous en persuader, nous avons placé la caméra face à un cops noir à $T_{CN} = 30°C$, et l'ensemble caméra et corps noir dans une chambre climatique. Nous avons fait évoluer la température de la chambre climatique (et donc de la caméra) et nous avons relevé les valeurs de 10 pixels choisies aléatoirement sur la matrice de microbolomètres. La Figure 64 montre les valeurs de ces 10 pixels sans que nous ayons corrigé la non-uniformité de l'image. La Figure 65 montre les valeurs de ces mêmes 10 pixels après avoir corrigé la non-uniformité de l'image grâce à une correction basée sur un obturateur interne. On

constante sur la Figure 64 que la non-uniformité de l'image est la différence entre les valeurs des pixels à une température de la caméra T_C donnée. On constate sur la Figure 65, qu'après avoir appliqué une correction de la non-uniformité *NU*, le niveau continu *NC* de l'image dépend toujours de la température de la caméra. C'est ce que nous appelons la dérive du niveau continu *NC* de l'image en fonction de la température de la caméra T_C.

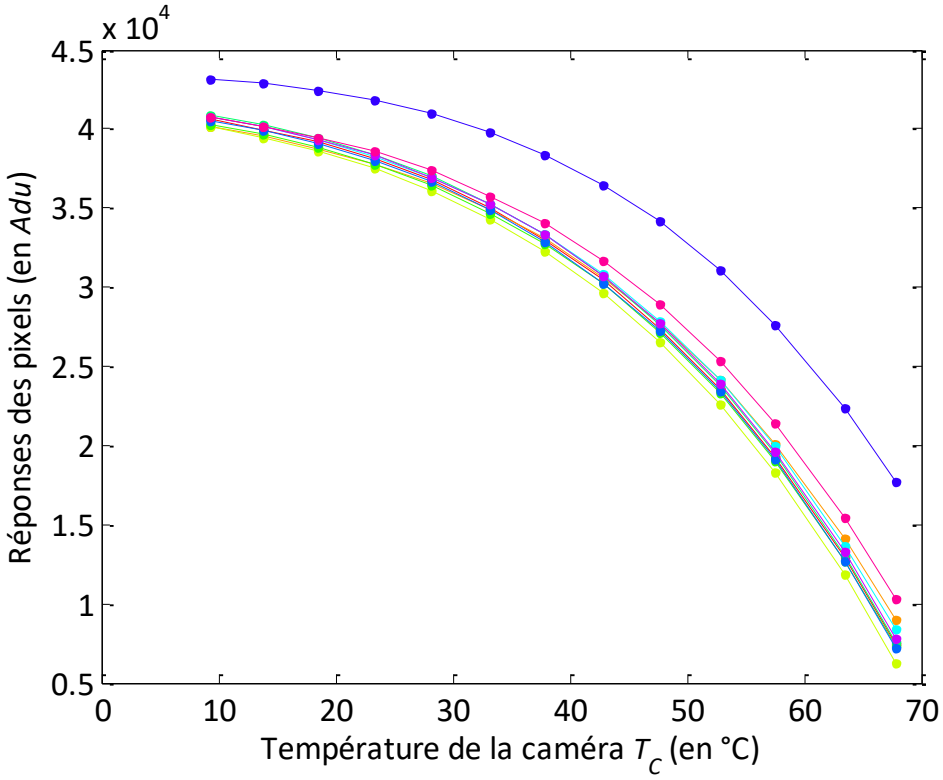


Figure 64. Valeurs (en *Adu*) de 10 pixels non corrigées de la non-uniformité *NU*, face à un corps noir à la température $T_{CN} = 30°C$ en fonction de la température de la caméra T_C.

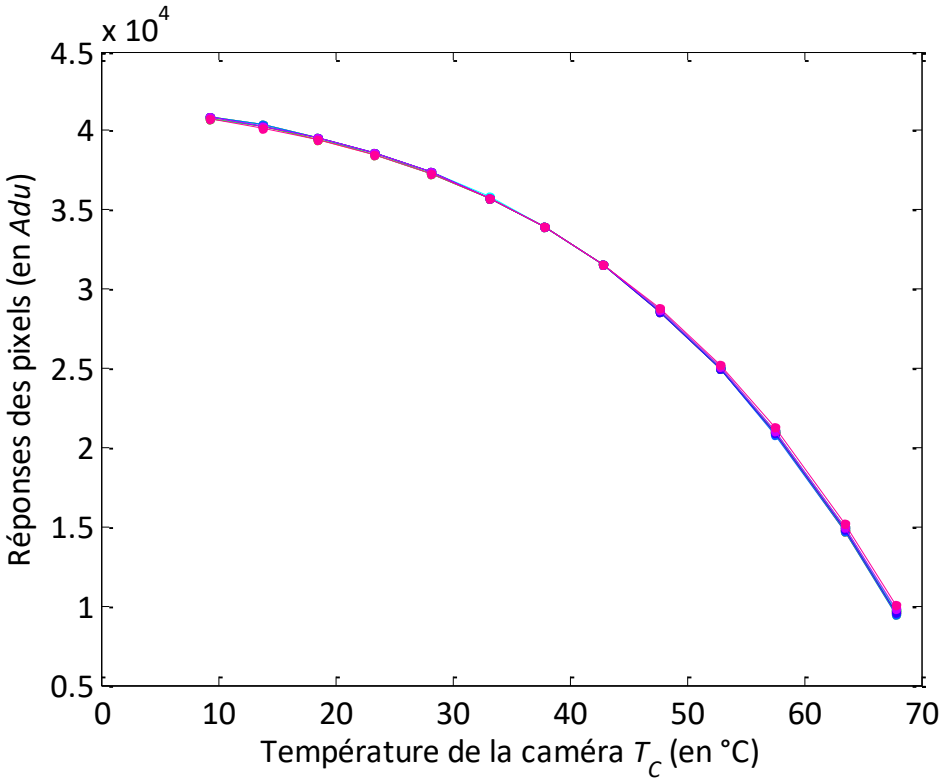


Figure 65. Valeurs (en *Adu*) de 10 pixels corrigées de la non-uniformité *NU* grâce à une méthode basée sur un obturateur interne. La caméra est face à un corps noir à la température $T_{CN} = 30°C$. La température de la caméra T_C évolue au cours de cette expérience.

Comme nous allons l'aborder au Chapitre 4, certains algorithmes de traitements d'images permettent d'estimer la pose 3D d'un objet (ou une partie) en minimisant une fonction de coût basée sur les moindres carrée pondérés entre l'image réelle et les images de synthèse issue d'une base que nous créons. Dans les images de synthèse, le visage adopte des poses spécifiques et labellisées. Ainsi être capable de détecter l'image de synthèse la plus proche de l'image réelle permet d'obtenir des informations sur la pose 3D (la création de la base d'images de synthèses sera détaillée au Chapitre 3). Plusieurs fonction de coût sont envisageables. Dans la plus simple, on considère que les valeurs des pixels d'une même

Comme nous allons l'aborder au Chapitre 5, certains algorithmes de traitement d'images permettent d'estimer la pose 3D d'un objet (orientation et position) en utilisant des dérivées (premières ou secondes) locales de l'image pour détecter et mettre en correspondance des points d'intérêt. Si i, j représentent les coordonnées d'un point d'intérêt, Y_{ij} est la valeur de l'image aux coordonnées i, j. Le gradient selon l'axe $\boldsymbol{i}$ est $Y_{i+1,j} - Y_{i-1,j}$ (cf. Figure 66).

En imagerie visible, ces gradients sont normalisés par rapport au niveau continu NC afin que l'algorithme soit robuste aux variations continues (ou basses fréquences) provoquées par des changements d'illumination. Le niveau continu *NC* d'une partie de l'image constituée de neuf pixels peut être représenté par la moyenne suivante :

$$NC = \frac{1}{9} \sum_{a=-1}^{1} \sum_{b=-1}^{1} Y_{i+a,j+b} \tag{2.43}$$

En imagerie thermique, particulièrement sur le visage, il y a peu d'informations (peu de points d'intérêt robustes). Les algorithmes traditionnellement développés pour le visible rencontrent des difficultés. Nous proposons d'utiliser le niveau continu *NC* du voisinage local des points d'intérêt, pour mieux les décrire comme dans la référence [101].

En effet, en imagerie thermique, seul le rayonnement propre de la scène est détecté par la caméra (les pixels voient également le rayonnement thermique du boîtier, qui est considéré comme un bruit de fond). De plus, le rayonnement thermique d'un visage est relativement constant car le corps humain est régulé en température. Cependant, la caméra utilisée est non-refroidie et comme cela est illustré sur la Figure 65, le niveau continu *NC* dépend de la température de la caméra T_C. En effet, le boîtier de la caméra va rayonner sur les pixels et l'image va être impactée. La conséquence de cela est qu'un point d'intérêt détecté sur l'image d'un visage (une zone de la joue, un coin d'œil…) possède un niveau qui dépend (i) du rayonnement propre du visage et (ii) de la température du boîtier. Utiliser le niveau continu *NC* de l'image pour décrire un point d'intérêt ne peut pas être implémenté sans prendre centaines précautions concernant cette dépendance à la température de la caméra T_C.

Au premier ordre, un offset global dépend de la température de la caméra (et d'autres températures come celle du plan focal, de l'optique…). Au second ordre, il est nécessaire de prendre en compte la *responsivité* car, comme nous l'avons vu dans la section 2.3, elle dépend de la température du plan focal.

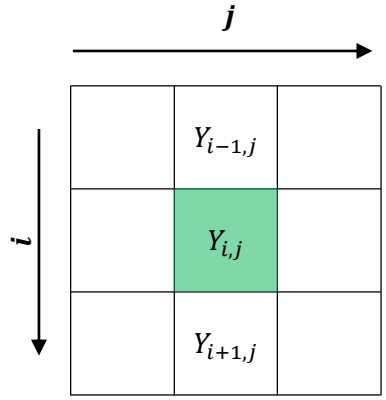


Figure 66. Valeur du pixel aux coordonnées i, j et de son voisinage.

Nous allons simplifier la problématique en considérant que le niveau continu *NC* et la *responsivité* dépendent uniquement de la température T_C, mesurée sur le boîtier de la caméra. La méthode qui va être

expliquée dans cette section consiste à rendre la réponse des pixels invariante à la température de la caméra en se basant sur des éléments de référence présents dans le champ de vue de la caméra.

2.6.2. Principe

On modélise les réponses des pixels comme suit :

$$\boldsymbol{Y} = \boldsymbol{\mathfrak{R}} \times \boldsymbol{F}_{scène} + \boldsymbol{\mathfrak{R}} \times \boldsymbol{F}_{boîtier} + \boldsymbol{D} \tag{2.44}$$

La matrice $\boldsymbol{Y}$ contient les valeurs des pixels de l'image (en Adu). Le terme $\boldsymbol{F}_{scène}$ est le flux rayonné par la scène et reçu par les pixels. Le terme $\boldsymbol{D}$ modélise une contribution (en *Adu*) qui dépend de la température du plan focal. Le terme $\boldsymbol{F}_{boîtier}$ désigne le flux reçu par les pixels et rayonné ou réfléchi par tous les éléments à l'intérieur de la caméra. La *responsivité* $\boldsymbol{\mathfrak{R}}$ correspond à la responsivité en (*Adu*/W). Les termes $\boldsymbol{\mathfrak{R}}$ et $\boldsymbol{U}$ dépendent essentiellement de la température du plan focal alors que $\boldsymbol{F}_{boîtier}$ dépend plutôt de la température du boîtier. En faisant l'hypothèse que la température du boîtier et celle du plan focal sont relativement proches, on peut regrouper les termes qui dépendent de la température de la caméra T_C : $\boldsymbol{U} = \boldsymbol{\mathfrak{R}} \times \boldsymbol{F}_{boîtier} + \boldsymbol{D}$, et on obtient :

$$\boldsymbol{Y} = \boldsymbol{\mathfrak{R}} \times \boldsymbol{F}_{scène} + \boldsymbol{U} \tag{2.45}$$

Ensuite en utilisant la *responsivité* à la radiance $\boldsymbol{\mathfrak{R}'}$ (en *Adu*.W^{-1}.m².sr) au lieu de la *responsivité* $\boldsymbol{\mathfrak{R}}$ en (en *Adu*.W^{-1}), on obtient :

$$\boldsymbol{Y} = \boldsymbol{\mathfrak{R}'} \times \boldsymbol{L}_{scène} + \boldsymbol{U} \tag{2.46}$$

Le terme $\boldsymbol{L}_{scène}$ correspond à la radiance de la scène. Dans l'équation (2.46), la *responsivité* en radiance $\boldsymbol{\mathfrak{R}'}$ dépend de la température de la caméra T_C, et nous connaissons cette dépendance (cf. équation (2.2)).

Avant de développer la méthode qui utilise un corps noir dans le champ de la caméra, citons quelques travaux scientifiques. Dans la référence [102] un nombre réduit d'acquisitions est nécessaire pour gérer la dépendance des réponses des pixels à la température de la caméra T_C. La *responsivité* $\boldsymbol{\mathfrak{R}'}$ et l'offset $\boldsymbol{U}$ sont modélisés par une fonction affine dépendant de la température de la caméra T_C :

$$\boldsymbol{Y} = (\boldsymbol{a} \times T_C + \boldsymbol{b}) \times \boldsymbol{L}_{scène} + (\boldsymbol{c} \times T_C + \boldsymbol{d}) \tag{2.47}$$

Les quatre tables de coefficient $\boldsymbol{a}$,$\boldsymbol{b}$,$\boldsymbol{c}$ et $\boldsymbol{d}$ sont à déterminer. Et, pour cela, il est nécessaire d'acquérir au minimum quatre cubes d'images du corps noir :

$$\boldsymbol{C}(T_{CN1}, T_{C1}),\ \boldsymbol{C}(T_{CN1}, T_{C2}),\ \boldsymbol{C}(T_{CN2}, T_{C1}) \text{ et } \boldsymbol{C}(T_{CN2}, T_{C2}),$$

Nous avons vu que ce qui prend du temps c'est d'attendre que la caméra soit stabilisée en température pour acquérir les cubes d'images. Dans cette méthode, une seule transition est nécessaire, ce qui réduit le temps d'étalonnage. Cependant les auteurs savent que cette modélisation atteint ses limites quand la température de la caméra varie dans un intervalle important. Dans ce cas, un polynôme d'ordre supérieur est nécessaire pour modéliser le second terme du membre de droite de l'équation (2.47). Dans ces travaux, l'offset est finalement modélisé par un polynôme d'ordre 3 et l'étalonnage de la caméra est réalisé en 18h. L'étalonnage

est ensuite testé pendant 24h dans lesquels la température du corps noir et la température de la caméra varient simultanément. Le but est d'estimer la température du corps noir T_{CN} malgré les variations de température de la caméra T_C. La température du corps noir T_{CN} varie de 10 à 50°C et la température de la caméra T_C varie de de 20 à 35°C. L'écart type de l'erreur d'estimation de la température T_{CN} du corps noir est de l'ordre de 0.3°C.

Dans la référence [103] (les auteurs sont les mêmes que ceux de la référence [102]), le niveau des pixels est rendu invariant à la température de la caméra en utilisant un *shutter* placé entre l'optique et la caméra. Des tables de coefficients établies par étalonnage en chambre climatique permettent de convertir l'image du *shutter* en une image d'un corps noir à la température de la caméra T_C, placé à l'extérieur de la caméra. Ces tables de coefficient prennent essentiellement en compte l'optique. Cette méthode est testée dans une expérience de 24h dans laquelle les températures du corps noir et de la caméra varient simultanément. Le corps noir recouvre le champ de vue *FOV* de la caméra. Le but est d'estimer au mieux sa température T_{CN} malgré les variations de température de la caméra T_C. La température du corps noir T_{CN} varie de 10 à 50°C et la température de la caméra T_C varie de de 20 à 35°C. Une erreur moyenne sur l'estimation de la température T_{CN} de l'ordre de 0.25°C et un écart type de l'ordre de 0.24°C sont obtenus.

Les méthodes citées jusqu'ici nécessitent une étape d'étalonnage très longue qui augmente le coût de la caméra. Il est toujours possible de se passer d'une étape d'étalonnage, mais en contrepartie, il est nécessaire qu'un certain nombre de pixels «voit » un corps de référence tel qu'un corps noir. On appelle ces pixels, pixels de « référence ». L'utilisation d'un objet de température et d'émissivité connues, peut également augmenter le prix du système ou tout simplement être inenvisageable dans certaines applications. Dans notre application, la caméra sera intégrée dans un habitacle d'automobile. Si il est envisageable d'introduire un tel objet dans l'habitacle sans un surcoût prohibitif, cette méthode a du sens.

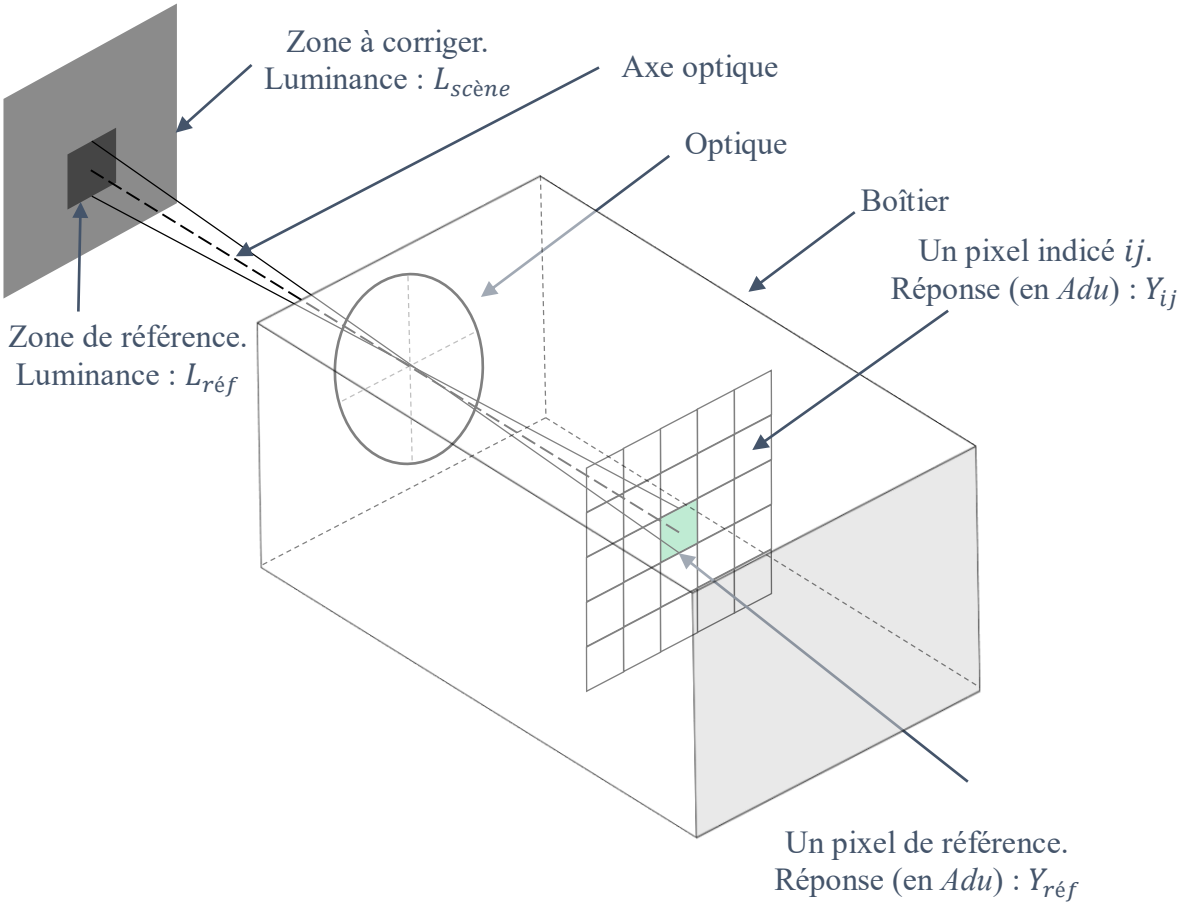


Figure 67. Schéma de principe de la compensation de la dérive des pixels grâce à une zone de référence. Sur cette représentation un seul pixel de référence est représenté.

La radiance de la scène $\boldsymbol{L_{scène}}$ est la grandeur physique qui nous intéresse. L'objectif de cette section est de l'extraire de l'équation (2.46). La stratégie consiste à estimer le terme $\boldsymbol{U}$ grâce à un élément extérieur étalonné. Nous considérons dans cette section qu'un corps noir est présent dans le champ de vue de la caméra. Certains pixels vont imager le corps noir. On appellera ces pixels, les pixels de référence (cf. Figure 67). La radiance de l'élément étalonné est notée $L_{réf}$. La réponse des pixels de référence s'exprime comme suit :

$$\boldsymbol{Y_{réf}} = \mathfrak{R}' \times L_{réf} + \boldsymbol{U_{réf}} \tag{2.48}$$

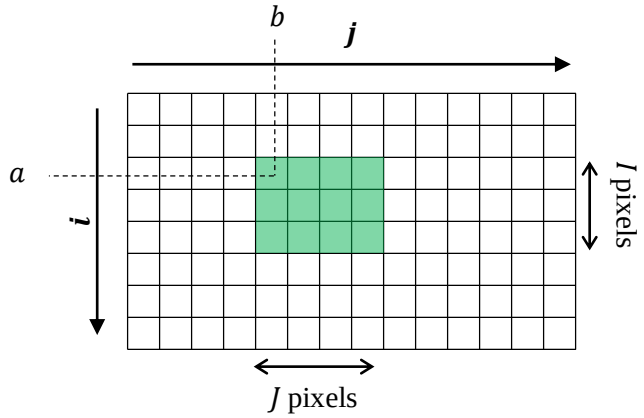


Figure 68. Position des pixels de référence.

L'équation (2.48) nous permet d'estimer la matrice $\boldsymbol{U_{réf}}$. Comme plusieurs pixels de référence sont utilisés, plusieurs estimations de $\boldsymbol{U_{réf}}$ sont effectuées (cf. Figure 68). Une moyenne de ces différentes estimations permettra de réduire les multiples sources de bruit (spatiales et temporelles) qui affectent cette estimation :

$$u_{réf} = \frac{1}{I \times J} \sum_{i=a}^{a+I} \sum_{j=b}^{b+J} U_{réf}(i,j) \tag{2.49}$$

La partie du signal correspondant au rayonnement du plan objet peut ainsi être estimée comme suit :

$$\boldsymbol{Y} - u_{réf} \tag{2.50}$$

Enfin, comme la *responsivité* $\mathfrak{R}'$ dépend de la température du plan focal, il est nécessaire de diviser l'expression (2.50) par $\mathfrak{R}'$ afin d'obtenir un terme indépendant de la température de l'environnement :

$$\frac{\boldsymbol{Y} - u_{réf}}{\mathfrak{R}'} \tag{2.51}$$

2.6.3. Expérience de dimensionnement

Nous avons souhaité simuler une telle approche. Voici le principe général en quatre points :

1. On fixe la température de la caméra T_C grâce à la température de la chambre climatique T_{CL}.
2. La caméra est face à un corps noir à $T_{CN} = 35°C$ et on enregistre un cube de 50 images.
3. On estime et on applique la compensation de la dérive des pixels en utilisant une zone de luminance $L_{réf}$ connue dans l'espace objet.
4. On fait varier la température de la caméra T_C en jouant sur la température de la chambre climatique T_{CL} et on recommence à partir de la seconde étape.

Concernant le matériel, l'idéal aurait été de posséder deux corps noirs. L'un aurait permis de créer la zone de référence, c'est-à-dire la zone de radiance connue $L_{réf}$, l'autre aurait permis d'avoir une scène de radiance $L_{scène}$.

Comme nous ne possédons qu'un seul corps noir, nous avons procédé différemment. Pour une température de la chambre climatique donnée, nous avons enregistré des images du corps noir à plusieurs températures T_{CN} : 30, 32, 35, 38 et 40°C. Nous faisons l'hypothèse que la température de la caméra T_C varie peu entre les différentes acquisitions des cubes d'images du corps noir. Ainsi pour corriger l'image du corps noir à 35°C, nous n'avons plus qu'à choisir les pixels de référence arbitrairement parmi les pixels de la matrice 480×640.

Tableau 5. Expérience d'étalonnage radiométrique grâce à un corps noir dans le champ de vue de la caméra.

Principe expérimental

Input : Les images de la caméra thermique représentent le flux rayonné par la scène et celui rayonné par le boîtier

Output : Les images de la caméra thermique représentent le flux rayonné par la scène.

For $T_{CL} = -2.5$ to 57.5°C

1. Acquérir les cubes d'images brutes du corps noir à $T_{CN} = 30, 32, 35, 38\ et\ 40°C$
2. Appliquer une correction de type NUC 2 points à tous les cubes
3. Choisir (arbitrairement) une température de corps noir pour jouer le rôle de source de luminance connue $L_{T_{réf}}$
4. Calculer $u_{réf}$
6. Estimer le flux rayonné par le corps noir à $T_{CN} = 35°C$:

$$\frac{\boldsymbol{Y} - u_{réf}}{\mathfrak{R}}$$

end

L'ensemble des acquisitions de cubes d'images de corps noir pour une température de la caméra donnée constitue un point de mesure (cf. Figure 69). Nous réalisons 13 points de mesure (de -2.5°C à 57.5°C par pas de 5°C.). Pour chaque point de mesure nous allons choisir arbitrairement (aléatoirement) la température de la zone de référence parmi celles disponibles : 30, 32, 35, 38 et 40°C. Pour tous les points de mesure, nous effectuons les étapes du Tableau 5 pour corriger les images du corps noir à 35°C. Après ceci, les pixels ne devraient pas varier en fonction des différents points de mesure.

Remarque : nous choisissons arbitrairement la température du corps noir de la zone de référence. Ainsi, en pratique une simple zone d'émissivité connue et de température mesurée (par un thermocouple par exemple) permet d'effectuer le même étalonnage.

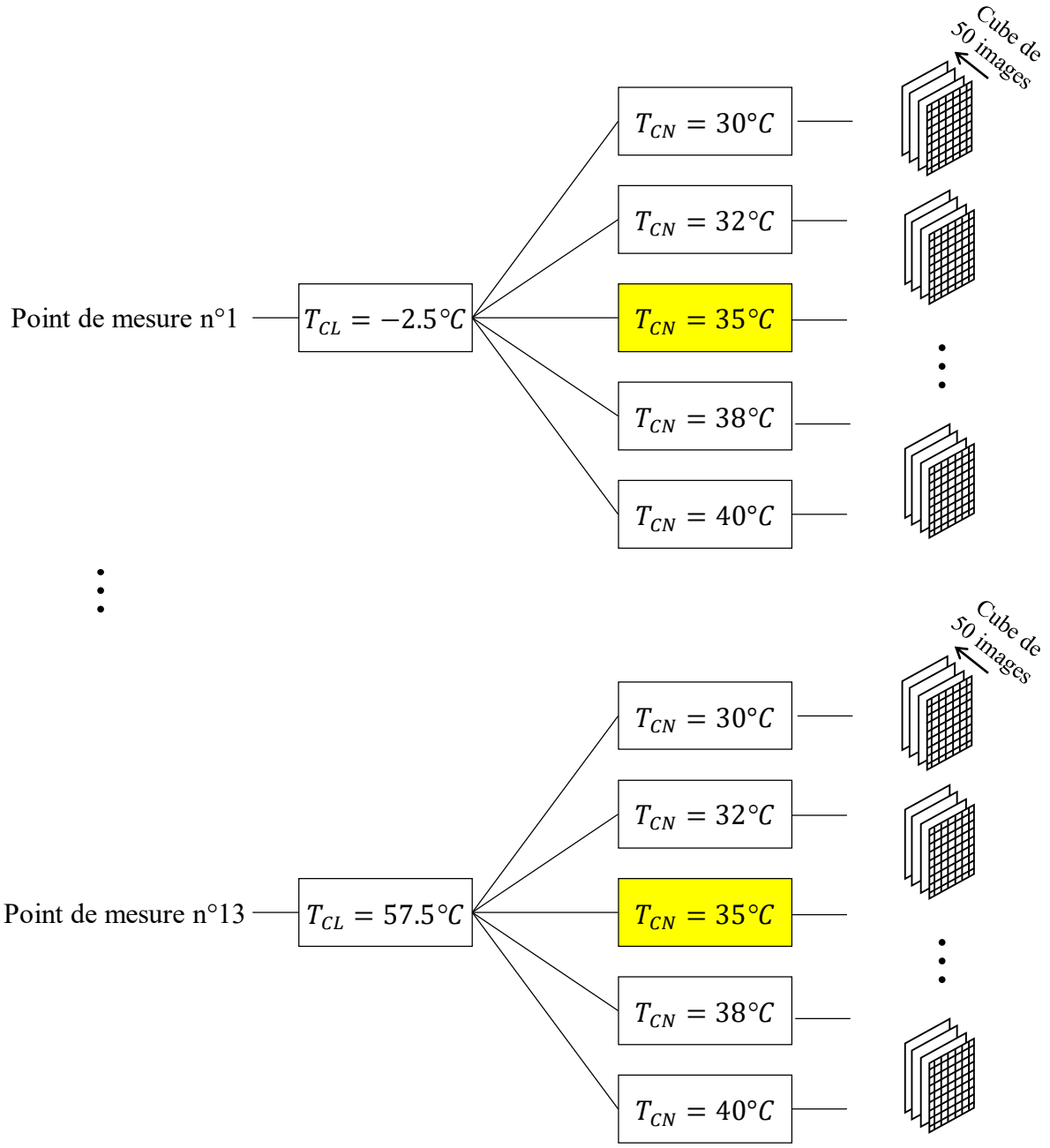


Figure 69. Données acquises pour l'expérience. Un point de mesure correspond à une température de la chambre climatique T_{CL}. Pour chaque point de mesure on enregistre 5 cubes d'images correspondant à 5 températures du corps noir T_{CN} (30, 32, 35, 38 et 40°C). Les cubes d'images du corps noir à 35°C (jaune sur le schéma) sont utilisés pour tester la compensation de la dérive des pixels. Les cubes utilisés pour la zone de référence sont choisis aléatoirement parmi les 5 températures de corps noir disponibles.

Tableau 6. Température T_C de la caméra pour les différents points de mesure pour l'expérience de compensation de la dérive. Toutes les températures sont en dégrées Celsius.

Points de mesure : température de la chambre climatique T_{CL} (en °C)	Température de la caméra T_C (en °C) (le corps noir est à T_{CN} =35°C)	Température de la zone de référence $T_{réf}$ (en °C) (choisie aléatoirement)	Température de la caméra T_C lors de l'acquisition de la zone de référence (le corps noir est à $T_{CN} = T_{réf}$)	Variation de la température de la caméra ΔT_C (en °C) entre l'instant d'acquisition de la zone à tester et l'instant d'acquisition de la zone de référence
-2.5	11.18	40	11.23	0.05
2.5	15.78	35	15.78	0
7.5	20.56	30	20.56	0
12.5	25.15	30	25.07	-0.08
17.5	29.89	32	29.85	-0.04
22.5	34.88	40	34.95	0.07
27.5	39.71	32	39.66	-0.05
32.5	44.71	40	44.79	0.08
37.5	49.71	32	49.67	-0.04
42.5	54.86	40	54.97	0.11
47.5	59.78	32	59.70	-0.08
52.5	65.51	30	65.42	-0.09
57.5	70.94	32	70.91	-0.03

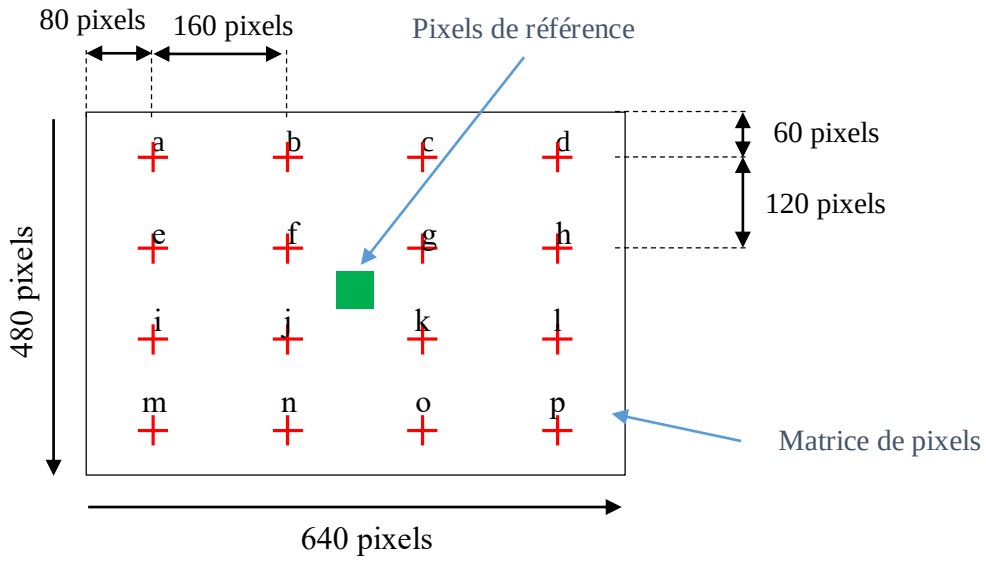


Figure 70. Position des pixels dont nous étudions les valeurs après compensation de la dérive. La zone verte correspond aux positions des 121 pixels de référence utilisés pour calculer la moyenne du bruit $\boldsymbol{B}$.

Nous considérons 121 pixels de référence. Les pixels de référence sont positionnés au centre de la matrice de microbolomètre de la ligne 235 à la ligne 245 et de la colonne 315 à 325, soit une fenêtre de 11×11 pixels (cf. carré vert sur Figure 70).

Pour quantifier l'effet de la compensation, nous étudions les valeurs de 16 pixels répartis comme cela est illustré sur la Figure 70. La Figure 71 illustre les valeurs des 16 pixels après compensation de la dérive et en fonction de la température de la caméra T_C.

La variation des réponses des pixels à un corps noir à 35°C a été largement réduite. Pour savoir si cette fluctuation est tolérable, je suis contraint d'annoncer deux résultats qui seront détaillés dans la suite de ce manuscrit :

(1) Lorsqu'un offset global entre l'image réelle d'un visage et les images de synthèse de ce même visage est inférieur à $\pm 320\ Adu$, l'estimation d'une partie de la pose par l'algorithme « global » est accélérée (cf. section 4.3.4).

(2) Lorsque les moyennes du voisinage locale des points d'intérêt sont stables dans un intervalle de $\pm 320\ Adu$, l'estimation de la pose par l'algorithme « local » est accelérée (cf. section 5.4.2.2).

Ces deux résultats indiquent que les variations de température de la caméra ne doivent pas générer des fluctuations des niveaux de l'image supérieures à un seuil exprimé en *Adu* ($\pm 320\ Adu$) par rapport à un état thermique de la caméra donnée (c'est-à-dire un état thermique pendant lequel nous avons acquis des images permettant de créer une base d'images synthétique utilisée pour l'estimation de la pose, cf. chapitre 3). Ce seuil a été évalué pour une température de caméra $T_C = 40.83°C$. Nous convertissons ce seuil en une radiance de la source grâce à la *responsivité* en radiance $\mathfrak{R}'$ à la température $T_C = 40.83°C$.

$$\langle \mathfrak{R}_{@40.83°C} \rangle = 191.3\ Adu.W^{-1}.m^2.sr$$

$$Seuil_{Adu} = 320\ Adu$$

$$Seuil_{W.m^{-2}.sr^{-1}} = 1.67\ W.m^{-2}.sr^{-1}$$

Le seuil sur le flux $Seuil_W$ va nous permettre de créer un critère de stabilité des pixels par rapport aux variations de la température de la caméra (cf. Tableau 7). Pour donner du sens à ce seuil, faisons remarquer que :

$$L_{31.9°C} - L_{30°C} = 1.65\ W.m^{-2}.sr^{-1}$$

Ramené en température, c'est comme si nous avions besoin d'une précision de ± 1.9°C sur l'estimation de la température de la scène. Ce qui semble tout à fait raisonnable car les étalonnages proposés avec les caméras commerciales ont ce type de performance.

Les variations des valeurs des pixels en fonction de la température de la caméra T_C sont illustrées sur la Figure 71.

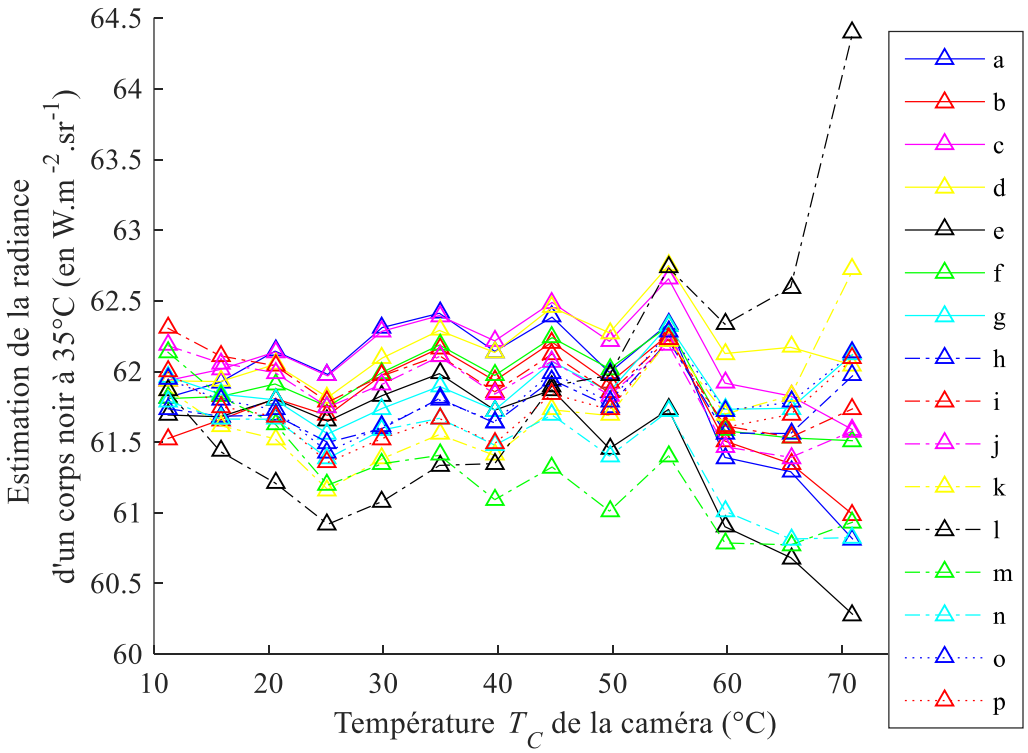


Figure 71. Valeurs des 16 pixels après compensation de la dérive de l'offset. La caméra est face à un corps noir à température constante $T_{CN} = 35°C$ pendant que la température de la caméra T_C évolue. Un capteur parfaitement compensé de la température de la caméra aurait une réponse plate.

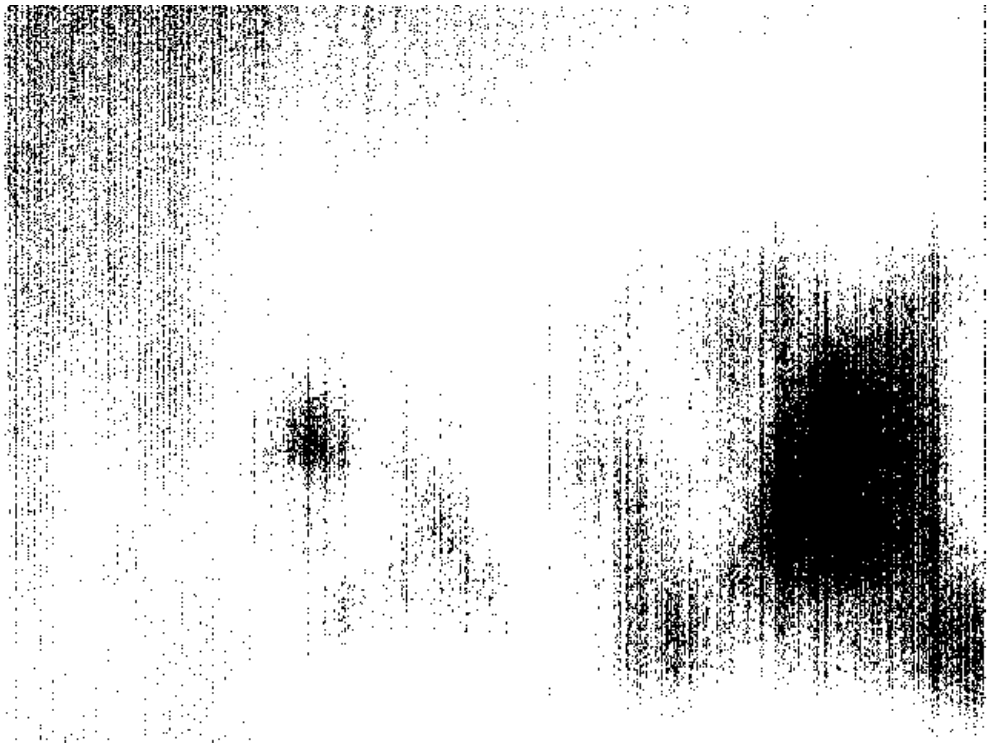


Figure 72. Carte des pixels respectant le critère de stabilité dans une plage de température $T_C \in [11.18{,}70.94°C]$.

Tableau 7. Critère de stabilité.

Critère de stabilité des pixels en fonction de la température de la caméra pour l'application d'estimation de la pose 3D du visage :

Soit $F_{ij@T_{CN}}(T_C)$ l'estimation du flux rayonné par un corps noir à la température T_{CN} stable temporellement et vu par le pixel i,j dans une plage de fonctionnement en température de la caméra comprise entre T_C^{min} et T_C^{max}.

Un pixel est considéré stable si l'écart crête à crête de $f_{ij@T_{CN}}(T_C)$ dans l'intervalle de température de fonctionnement de la caméra $[T_C^{min}, T_C^{max}]$ est inférieur à $Seuil_W$

$$\max\left(F_{ij\,@T_{CN}}(T_C)\right) - \min\left(F_{ij\,@T_{CN}}(T_C)\right) < Seuil_W$$

La Figure 72 représente la localisation des pixels qui respectent le critère de stabilité dans la gamme de température de fonctionnement $T_C \in [11.18{,}70.94°C]$. Sur cette carte, un pixel indicé ij,

- vaut 1 si $\max\left(f_{ij\,@T_{CN}}(T_C)\right) - \min\left(f_{ij\,@T_{CN}}(T_C)\right) < Seuil_W$,
- 0 sinon.

Dans la gamme de températures $T_C \in [11.18{,}70.94°C]$, 89.1% des pixels respectent le critère de stabilité. Si nous réduisons la gamme de températures de la caméra à $T_C \in [25.2\,;\,54.9°C]$, 99.8% des pixels respectent le critère de stabilité.

On remarque sur la Figure 71 qu'au-delà de $T_C = 60°C$ cette méthode ne fonctionne plus. L'hypothèse que la température de la caméra est homogène (la température du boîtier est proche de celle du plan focal) n'est probablement plus respectée. La Figure 72 nous renseigne sur la localisation des zones où cette hypothèse est la moins bien respectée.

Cette section montre qu'il est possible, grâce à un élément de référence placé dans le champ de vue de la caméra, d'effectuer un étalonnage radiométrique. Celui-ci permettra d'accélérer les algorithmes de traitement des images.

Chapitre 3. Réalisation d'un maillage 3D texturé et d'une base d'images de synthèse

3.1. Introduction

Dans la limite de nos connaissances, il n'a pas encore été montré dans la littérature scientifique qu'il était possible d'utiliser un système d'imagerie thermique non-refroidie monoculaire pour estimer la pose du visage avec une précision importante. Dans ce chapitre, nous introduirons un certain nombre de principes utilisés en traitement d'image dans le domaine spécifique de la reconstruction 3D. Puis, nous discuterons en détails de l'élément qui permet de gagner en précision : le maillage 3D texturé du visage.

Dans la section 3.2 les principes géométriques de formation d'une image sur la rétine d'une caméra à l'aide du modèle caméra *sténopé* (également appelé *pinhole* dans la littérature anglophone) seront rappelés. Le calibrage géométrique de caméra utilisée dans ce projet sera également présenté. Dans la section 3.4 nous expliquerons pourquoi il est nécessaire d'utiliser un modèle géométrique 3D du visage. Dans la section 3.5 deux maillages 3D du visage seront présentés. Ces maillages 3D sont rigides et nous leur imposons un centre de rotation au milieu de l'axe interaural (entre les oreilles). Ce choix sera détaillé. Nous expliquerons également comment plaquer sur ces maillages 3D la texture issue des images thermiques réelles. La base d'images synthétiques créée à partir du maillage 3D texturé sera aussi présentée.

3.2. Le modèle caméra sténopé

Cette section a pour objectif de rappeler les notions qui permettent de comprendre la formation d'une image sur la rétine d'une caméra et surtout d'introduire les notations utilisées dans ce manuscrit. Les points (2D et 3D), les vecteurs et les matrices sont exprimés en gras. La distinction entre majuscule et minuscule ne désigne rien en particulier. La notation $\boldsymbol{A}^T$ représente la transposée de la matrice $\boldsymbol{A}$. Toutes les autres lettres pouvant apparaitre en exposant indiquent le repère dans lequel est exprimé l'élément en question. Une lettre en indice permet d'identifier l'élément en question.

Commençons d'abord par définir le terme focale dans le domaine de l'optique et dans celui de la vision par ordinateur.

La distance *f* dans le domaine de la vision par ordinateur : la distance *f* est une valeur positive qui désigne toujours la distance entre le centre de projection et le plan image. Dans le cas d'une conjugaison qui n'est pas *infinie-foyer*, *f* n'est pas la distance focale et son utilisation est abusive au sens de l'optique. Le terme de tirage serait plus approprié au sens de l'optique (cf. schéma a) de la Figure 73).

La focale dans le domaine de l'optique : dans le cas d'une lentille mince dans l'air, la distance focale image est la distance entre le centre de la lentille et le foyer image (cf. schéma b) de la Figure 73). La distance focale image est une valeur algébrique qui dépend de l'indice des matériaux traversés ainsi que des rayons de courbures. Le foyer image correspond à la position de l'image net d'un objet situé à l'infinie

(c'est-à-dire lorsque la valeur absolue de la distance entre l'objet et la caméra est grande devant la valeur absolue de la distance focale).

Dans ce chapitre qui traite majoritairement d'aspects du domaine de la vision par ordinateur, la focale désignera toujours la distance entre le centre de l'optique et la rétine, même si la conjugaison n'est pas *infini-foyer*.

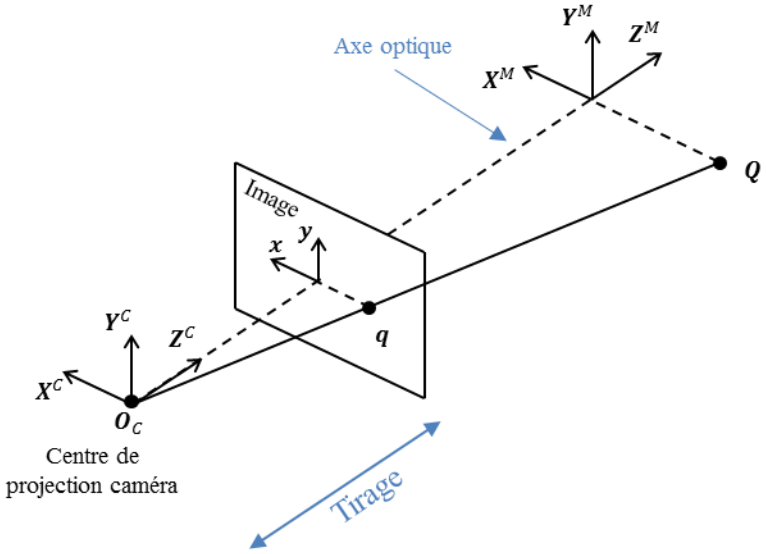


Schéma a) : modèle *sténopé*

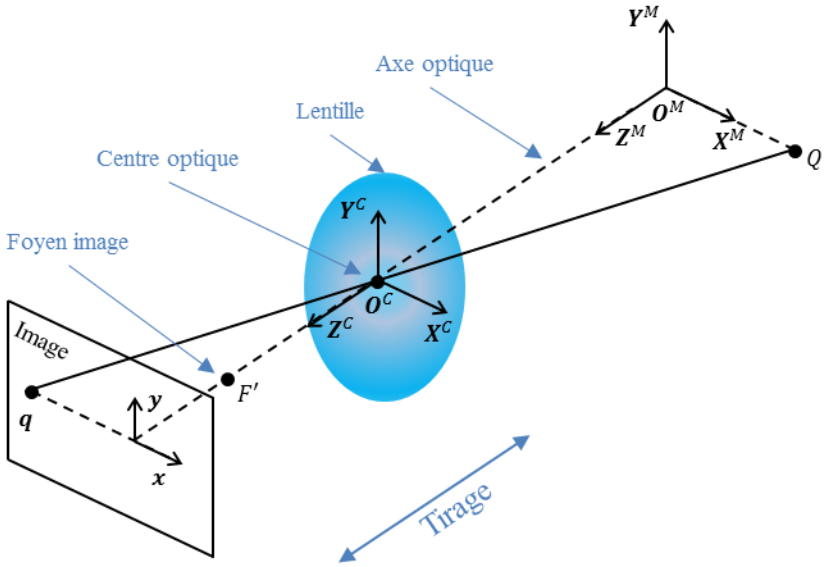


Schéma b) : modèle *lentille mince*

Figure 73. Modèle « sténopé » (domaine du traitement d'images) et modèle « lentille mince » (domaine de l'optique). Schéma a) représente le modèle « stenopé » utilisé dans le domaine de la vision par ordinateur. La focale au sens de l'optique n'est généralement pas représentée et on fait l'approximation d'une conjugaison *infini-foyer*. Ainsi le tirage est souvent désigné par le terme focal par abus de langage. Le schéma b) représente le modèle « lentille mince » dans le domaine de l'optique. La distance entre le centre de la lentille O^C et le foyer image F' est la distance focale.

Remarquons ensuite que le centre de projection dans le modèle *sténopé* correspond au centre de l'optique dans le modèle d'une lentille mince. De plus, dans le modèle *sténopé*, le plan image est positionné entre l'objet et le centre de projection alors que dans le modèle de la lentille mince, le système est 'déplié' et c'est la lentille qui est entre le plan objet et le plan image. Le grandissement transversal défini la taille de l'image sur celle de l'objet. Dans ces deux modèles, le grandissement a la même valeur absolue mais un signe opposé.

Dans ce chapitre nous utilisons le modèle *sténopé*, nous allons donc le présenter plus en détail. Dans ce modèle, une caméra est définie par un repère orthonormé noté $(\boldsymbol{O^C}, \boldsymbol{X^C}, \boldsymbol{Y^C}, \boldsymbol{Z^C})$ où $\boldsymbol{O^C} = [0,0,0]^T$, $\boldsymbol{X^C} = [1,0,0]^T$, $\boldsymbol{Y^C} = [0,1,0]^T$ et $\boldsymbol{Z^C} = [0,0,1]^T$. $\boldsymbol{O^C}$ est le centre de projection de la caméra (cf. Figure 74). La direction du vecteur $\boldsymbol{Z^C}$ défini l'axe optique. Le plan image est défini par le repère 2D orthonormé $(\boldsymbol{o}, \boldsymbol{x}, \boldsymbol{y})$. Le point noté $\boldsymbol{o}$ est appelé le point principale de l'image (ou centre de l'image). Il s'agit de l'intersection de l'axe optique avec le plan image. La distance *f*, appelée focale par abus de langage comme signalé précédemment, correspond à la distance entre le centre de projection et le plan image.

$$f = \overline{\boldsymbol{O^C o}} \tag{3.1}$$

Pour illustrer les notations que nous utilisons, définissons les coordonnés 3D d'un point objet $\boldsymbol{Q}$ dans le repère monde :

$$\boldsymbol{Q}^M = \begin{bmatrix} X^M \\ Y^M \\ Z^M \end{bmatrix} \tag{3.2}$$

Lorsque l'on souhaite identifier un point compris dans un ensemble, on utilise la notation indicielle. Par exemple le point $\boldsymbol{Q_i}$ dans le repère monde s'exprime comme suit :

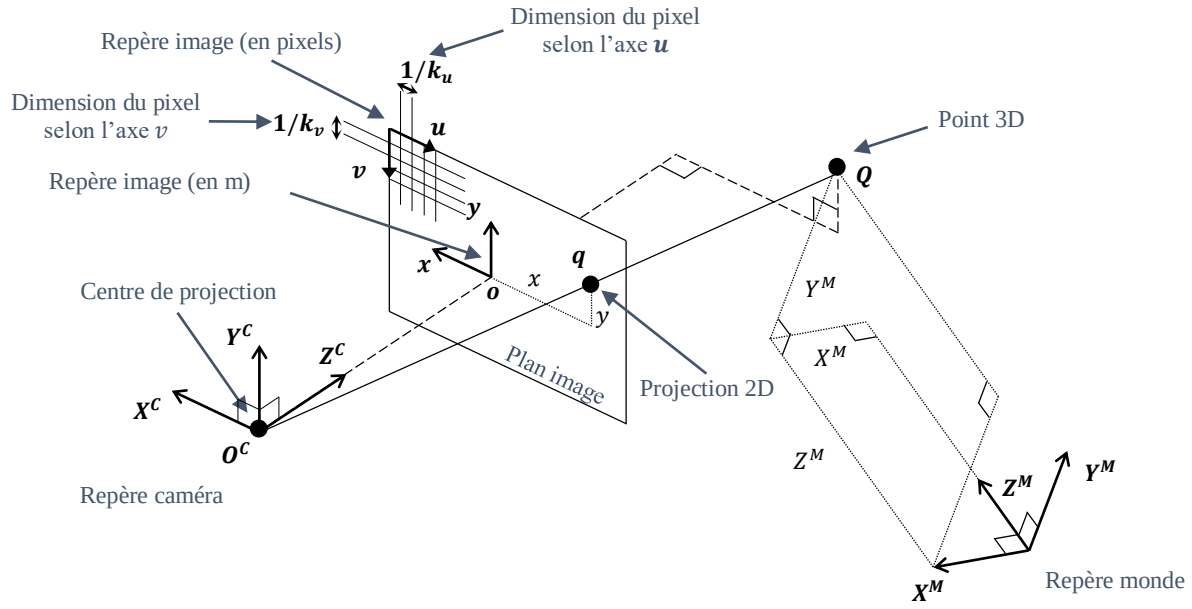


Figure 74. Le modèle sténopé : schéma de principe et notations.

$$\boldsymbol{Q_i^M} = \begin{bmatrix} X_i^M \\ Y_i^M \\ Z_i^M \end{bmatrix} \tag{3.3}$$

Le point de vue, appelé également la pose, est la position et l'orientation du repère caméra par rapport au repère monde. La pose est ainsi définie par 6 paramètres : les 3 coordonnées d'un vecteur de translation et les 3 angles d'une matrice de rotation. La projection 2D sur le plan image d'un point 3D est noté $\boldsymbol{q}$. Elle dépend de la pose et de la focale de la caméra. La section suivante introduit le modèle utilisé pour décrire la formation de l'image.

Remarque : Considérer que le repère de la caméra est en mouvement relatif et le repère monde est fixe est équivalent à considérer que le repère de la caméra est fixe et le repère monde en mouvement relatif.

3.2.1. Modèle de projection (paramètres intrinsèques)

L'objectif de cette section est d'expliquer le lien entre les coordonnées 3D (en m) d'un point exprimé dans le repère caméra et les coordonnées 2D (en pixels) de sa projection exprimée dans le plan image de la caméra. Le point $\boldsymbol{Q}$ exprimé dans le repère camera est noté $\boldsymbol{Q^C}$. Il possède les coordonnées suivantes :

$$\boldsymbol{Q^C} = \begin{bmatrix} X^C \\ Y^C \\ Z^C \end{bmatrix} \tag{3.4}$$

La projection 2D est notée q :

$$\boldsymbol{q} = \begin{bmatrix} x \\ y \end{bmatrix} \tag{3.5}$$

Dans la suite du manuscrit, il arrivera parfois que le repère image soit également précisé pour la projection. On pourra par exemple trouver la notation q^A qui sous-entend qu'on parle de la projection associée à la caméra A.

Grâce aux triangles semblables $(O_C QH)$ et $(O_C qH)$, on peut déduire les relations suivantes :

$$x = f\frac{X^C}{Z^C} \tag{3.6}$$

$$y = f\frac{Y^C}{Z^C} \tag{3.7}$$

Sous forme matricielle on obtient la relation suivante :

$$\boldsymbol{q} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \sim \begin{bmatrix} f.X^C \\ f.Y^C \\ Z^C \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X^C \\ Y^C \\ Z^C \end{bmatrix} \tag{3.8}$$

Le symbole $\sim$ exprime la proportionnalité et l'idée qu'une dimension spatiale est perdue par effet de projection. En d'autres termes, tout point $\boldsymbol{Q_j}$ aligné avec un point $\boldsymbol{Q_i}$ et le centre de projection $\boldsymbol{O^C}$ sera projeté au même endroit que $\boldsymbol{Q_i}$. Dans l'équation (3.8) les coordonnées de $\boldsymbol{q}$ sont normalisées par rapport à la troisième composante. On parle de coordonnées image homogènes.

Il est ensuite nécessaire de modéliser la discrétisation liée à l'échantillonnage des pixels. Le passage du repère image à des coordonnées $[u, v]^T$ en pixels s'exprime de la manière suivante :

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} k_u & 0 & 0 \\ 0 & k_v & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & x_0 \\ 0 & 1 & y_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.9}$$

Les densités de pixels (c'est-à-dire le nombre de pixels par unité métrique) dans les directions respectives $\boldsymbol{u}$ et $\boldsymbol{v}$ (cf. Figure 74) sont notées k_u et k_v. Les coordonnées du centre optique (en m) sont notées x_0 et y_0.

La combinaison des équations (3.8) et (3.9) permet de modéliser le passage du repère caméra au repère image :

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} k_u f & 0 & k_u x_0 \\ 0 & k_v f & k_v y_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X^C \\ Y^C \\ Z^C \end{bmatrix} \tag{3.10}$$

La matrice 3 × 3 noté $\boldsymbol{K}$ est souvent utilisée dans la littérature pour regrouper les paramètres intrinsèques de la caméra :

$$\boldsymbol{K} = \begin{bmatrix} k_u f & 0 & k_u x_0 \\ 0 & k_v f & k_v y_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.11}$$

Lorsque la matrice du détecteur est composée de pixels de formes carrées, on peut faire l'approximation que $k_u = k_v$. Dans ce cas la matrice peut s'écrire comme suit :

$$\boldsymbol{K} = \begin{bmatrix} \alpha_0 & 0 & u_0 \\ 0 & \alpha_0 & v_0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{array}{ll} avec & \alpha_0 = k_u f, \\ & u_0 = k_u x_0, \\ et & v_0 = k_u y_0. \end{array} \tag{3.12}$$

Le terme α_0 représente la focale en nombre de pixels. De même, u_0 et v_0 représente les coordonnées du centre optique en pixels.

3.2.2. La pose de l'objet par rapport à la caméra (paramètres extrinsèques)

Dans un cadre plus général, les coordonnées 3D du point $\boldsymbol{Q}$ sont exprimées dans un repère monde ($\boldsymbol{Q}$ est noté $\boldsymbol{Q^M}$ dans le repère monde) qui n'est pas aligné avec le repère de la caméra comme cela est illustré sur la Figure 74 ($\boldsymbol{Q}$ est noté $\boldsymbol{Q^C}$ dans le repère caméra). La rotation $\boldsymbol{R}$ ainsi que la translation $\boldsymbol{t}$ qui lient ces deux repères ne sont généralement pas connues. Comme déjà évoqué plus haut, la combinaison de ces deux transformations est la pose. On parle également des paramètres extrinsèques de la caméra par opposition aux paramètres intrinsèques. Ainsi le point $\boldsymbol{Q^M}$ exprimé dans le repère monde peut également être défini dans le repère caméra grâce aux paramètres extrinsèques :

$$\begin{bmatrix} X^C \\ Y^C \\ Z^C \end{bmatrix} = \boldsymbol{R} \begin{bmatrix} X^M \\ Y^M \\ Z^M \end{bmatrix} + \boldsymbol{t} \tag{3.13}$$

La matrice de rotation $\boldsymbol{R}$ est définie par les trois angles d'Euler. Elle est le produit de trois matrices de rotation :

$$\boldsymbol{R} = Rot(Z,\psi)Rot(Y,\varphi)Rot(X,\theta) \tag{3.14}$$

Les coordonnées homogènes (ajout d'une quatrième composante qui vaut 1) permettent d'écrire le changement de repère sous forme d'une multiplication matricielle :

$$\begin{bmatrix} X^C \\ Y^C \\ Z^C \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ 0_{1\times 3} & 1 \end{bmatrix} \begin{bmatrix} X^M \\ Y^M \\ Z^M \\ 1 \end{bmatrix} \tag{3.15}$$

La matrice $0_{1\times 3}$ d'une ligne et de trois colonnes contient uniquement des zéros $0_{1\times 3} = [0,0,0]$. Des notations encore plus condensées négligent les éléments utiles à l'écriture en cordonnées homogènes :

$$\begin{bmatrix} X^C \\ Y^C \\ Z^C \\ 1 \end{bmatrix} = [\boldsymbol{R}|\boldsymbol{t}] \begin{bmatrix} X^M \\ Y^M \\ Z^M \\ 1 \end{bmatrix} \tag{3.16}$$

3.2.3. Le modèle *sténopé* complet

La combinaison de la projection (équation (3.10)) et du changement de repère (équation (3.16)) permet de modéliser complétement le processus de formation d'une image d'un objet 3D sur le plan image d'une caméra. L'utilisation des coordonnées homogènes nécessite d'ajouter une colonne de 0 à la matrice $\boldsymbol{K}$. Ainsi la matrice $\boldsymbol{K_{3\times 4}}$ est définie comme suit :

$$\boldsymbol{K_{3\times 4}} = \begin{bmatrix} \alpha_0 & 0 & u_0 & 0 \\ 0 & \alpha_0 & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{3.17}$$

Finalement le processus de formation d'image s'écrit :

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \boldsymbol{K_{3\times 4}} \begin{bmatrix} \boldsymbol{R_{3\times 3}} & \boldsymbol{t_{3\times 1}} \\ 0_{1\times 3} & 1 \end{bmatrix} \begin{bmatrix} X^M \\ Y^M \\ Z^M \\ 1 \end{bmatrix} \tag{3.18}$$

On utilisera dans ce manuscrit la matrice $\boldsymbol{P}$ de taille 3×4 pour modéliser le processus complet de projection. $\boldsymbol{P}$ est appellée la matrice de projection et s'exprime comme suit :

$$\boldsymbol{P} = \boldsymbol{K_{3\times 4}} \begin{bmatrix} \boldsymbol{R_{3\times 3}} & \boldsymbol{t_{3\times 1}} \\ \boldsymbol{0_{1\times 3}} & \boldsymbol{1} \end{bmatrix} \tag{3.19}$$

Dans le domaine de la vision par ordinateur l'estimation de la pose est très populaire. Ses applications sont la robotique, la réalité augmenté… On distingue deux types de problématique. Il est possible d'estimer la pose entre :

- deux images successives (cf. section 3.3),
- un objet 3D connu et une caméra (cf. chapitre 5).

Dans les deux cas, ce sont les paramètres extrinsèques qui sont utiles à l'application. La connaissance des paramètres intrinsèques est une étape simplement indispensable. Il est souvent acceptable de considérer une étape de calibrage géométrique usine pour définir $\boldsymbol{K}$.

3.3. Calibrage géométrique

Le calibrage géométrique (aussi appelé étalonnage géométrique ou encore *calibration* dans la littérature anglophone) permet d'estimer les paramètres intrinsèques de la caméra : la focale, la position du centre optique dans le plan image et les coefficients de distorsion (radiale).

Nous avons utilisé la fonction *estimateCameraParameters* de Matlab 2015b pour déterminer les paramètres intrinsèques de la caméra *Xenics Gobi 640 CL*. Cet algorithme utilise plusieurs images d'une mire plane acquise de différents points de vue. Dans un premier temps les paramètres intrinsèques et extrinsèques de la caméra sont estimés en considérant que la distorsion de la lentille est nulle [104]. Dans un second temps, l'estimation de ces paramètres est affinée et la distorsion est estimée. Une méthode de minimisation des moindres carrés non-linéaire basée sur l'algorithme Levenberg–Marquardt permet d'effectuer cette seconde étape [105]. La minimisation est initialisée avec une distorsion nulle et les paramètres intrinsèques et extrinsèques de la première étape.

3.3.1.1. Mire à damier

Une mire à damier est souvent utilisée en imagerie visible. En imagerie thermique on observe la radiation propre de la surface des objets. Il n'est pas aisé de créer une mire en contrôlant spatialement la température de la surface de l'objet. Généralement on préfère utiliser des mires qui fonctionnent par transmission [106,30]. Nous avons créé une mire à damier qui fonctionne par transmission. Pour cela un motif à damier a été imprimé sur un film plastique (cf. Figure 75). Le motif est composé de 13 × 14 carrés noirs ou blancs, de 4.5 mm. Cette mire est ensuite fixée devant le corps noir (cf. Figure 76). Nous réglons la température du corps noir à 40°C et nous enregistrons 16 images grâce à la caméra thermique et avec des points de vue différents (cf. Figure 77).

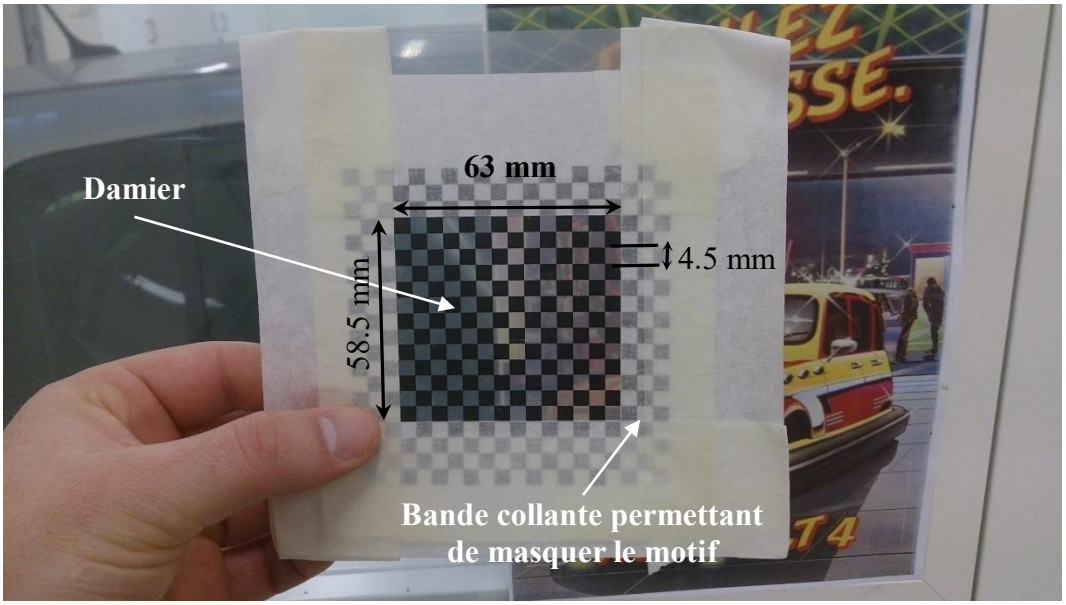


Figure 75. Mire à damier par transmission pour le calibrage géométrique de la caméra thermique.

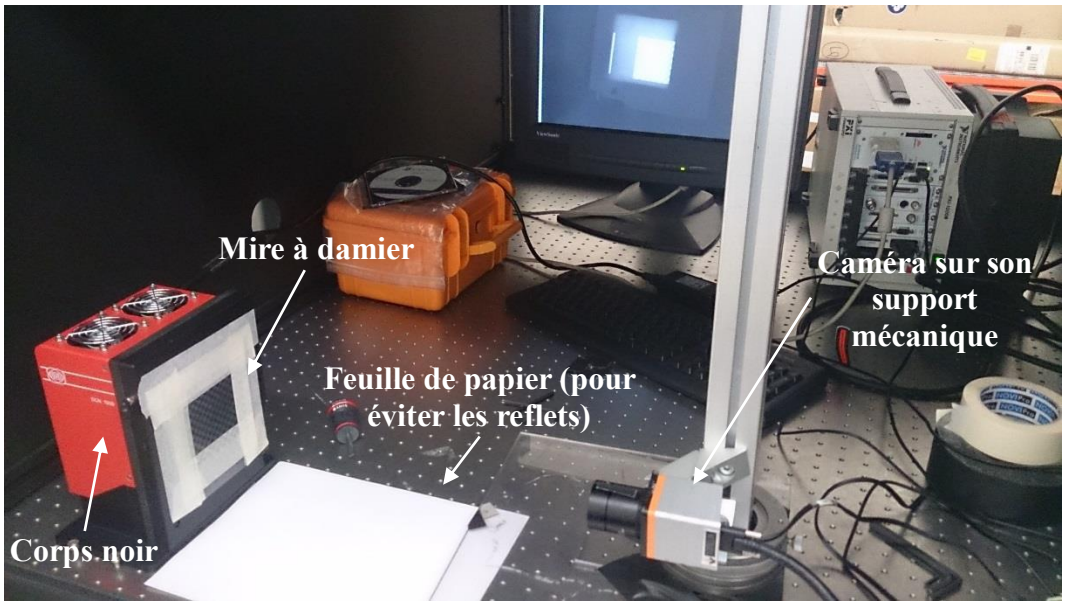


Figure 76. Disposition du matériel pour le calibrage géométrique de la caméra thermique à l'aide d'une mire à damier.

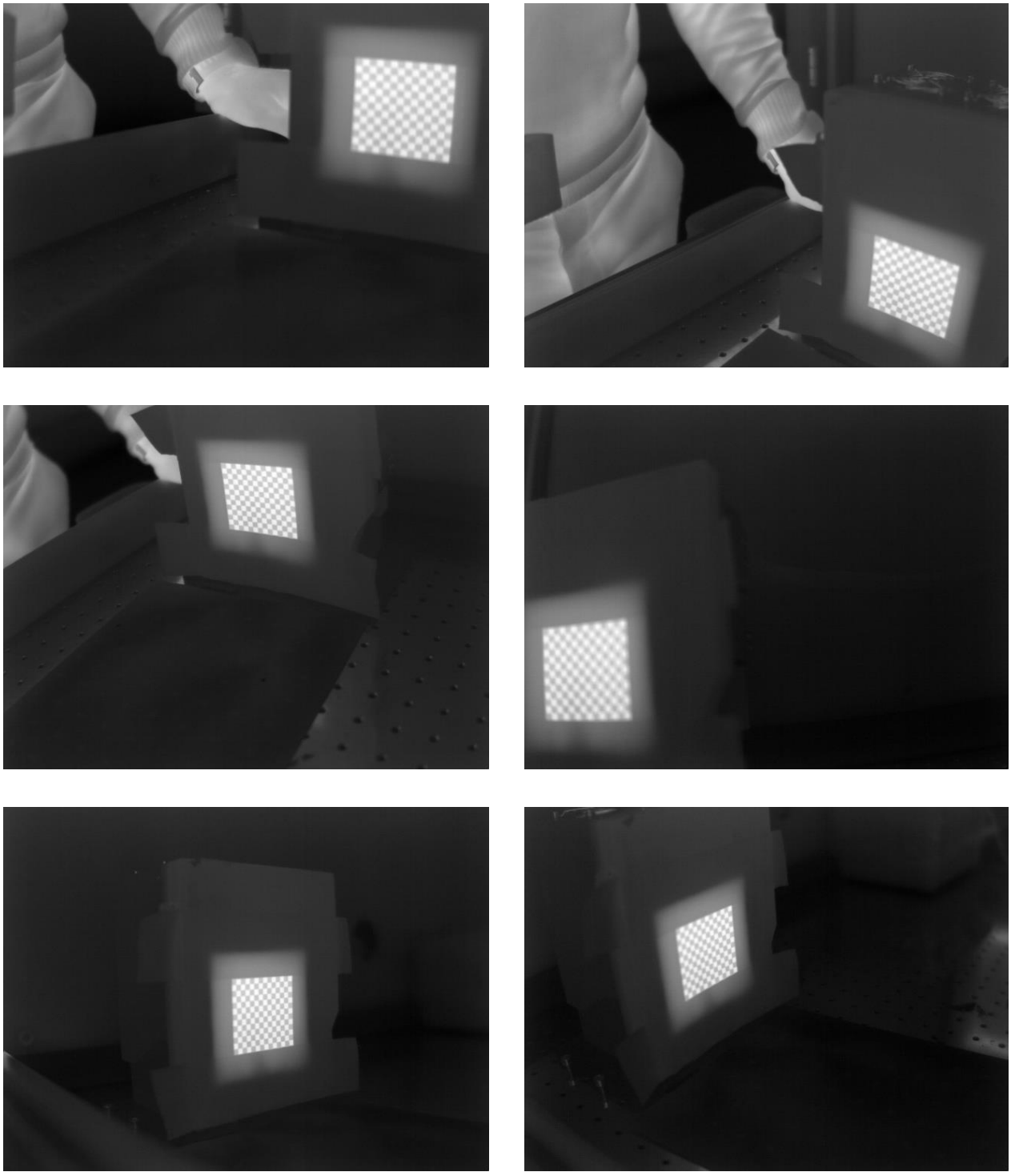


Figure 77. Six des seize images utilisées pour le calibrage géométriques.

Nous utilisons ensuite la fonction *detectCheckerBoardPoints* de Matlab 2015b pour détecter les coins de la mire à damier (cf. Figure 78). A partir des points détectés sur les images de la mire à damier et des dimensions connues du motif à damier, la fonction *estimateCameraParameters* estime les paramètres intrinsèques de la caméra. La focale (le terme focale est utilisé par abus de langage car la conjugaison n'est pas de type *infinie-foyer*) de la caméra est évaluée à 785 pixels. Sachant que la taille du pixel vaut 17 µm, cela équivaut à une distance image de 13.35 mm. Pour rappel, l'objectif possède une distance focale optique (c'est-à-dire la véritable distance focale) de 12 mm. En utilisant les formules de conjugaison avec origine au centre, le conjugué du plan image est un plan objet situé 108 mm en avant de l'objectif. Or la mise au point a été réalisée pour un objet à ~55 cm en avant de l'objectif.

L'estimation de la position du centre optique en pixel dans le repère image est $[u, v]^T = [262{,}324]^T$. Pour rappel, la caméra possède une matrice de pixel au format $640 \times 480\ pixels$. Les coefficients de distorsion sont $k_1 = -0.47$ et $k_2 = 0.50$.

L'erreur de reprojection est un critère usuellement utilisé pour évaluer la précision de l'estimation des paramètres intrinsèques. Elle mesure la distance entre un point détecté sur une image et la projection d'un point 3D grâce aux paramètres intrinsèques et extrinsèques estimés. L'erreur de reprojection donnée pour une image est une moyenne des erreurs de reprojection de tous les points de la mire (cf. Figure 79, pour chaque image est donnée l'erreur de reprojection moyenne). L'erreur de reprojection globale est une moyenne sur toutes les images utilisées pour le calibrage. Elle vaut en moyenne 0.22 pixels pour ce calibrage.

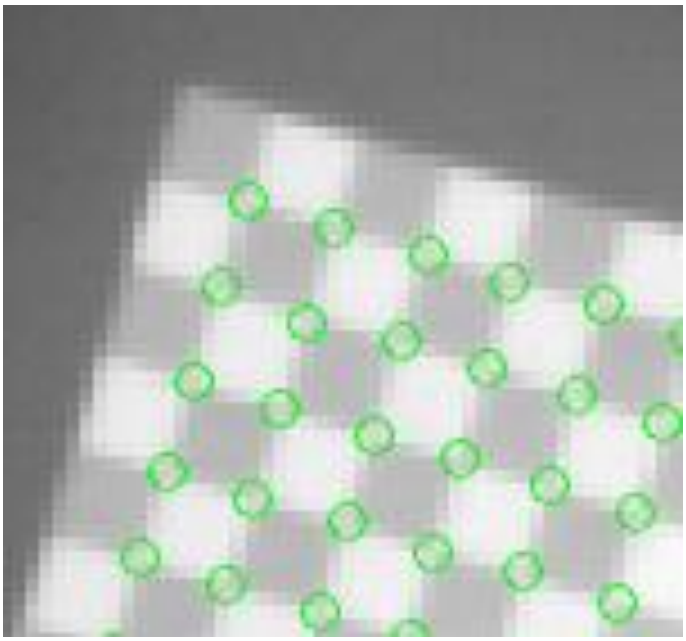


Figure 78. Détection des amers sur la mire à damier grâce à la fonction *detectCheckerBoardPoints* de Matlab 2015b.

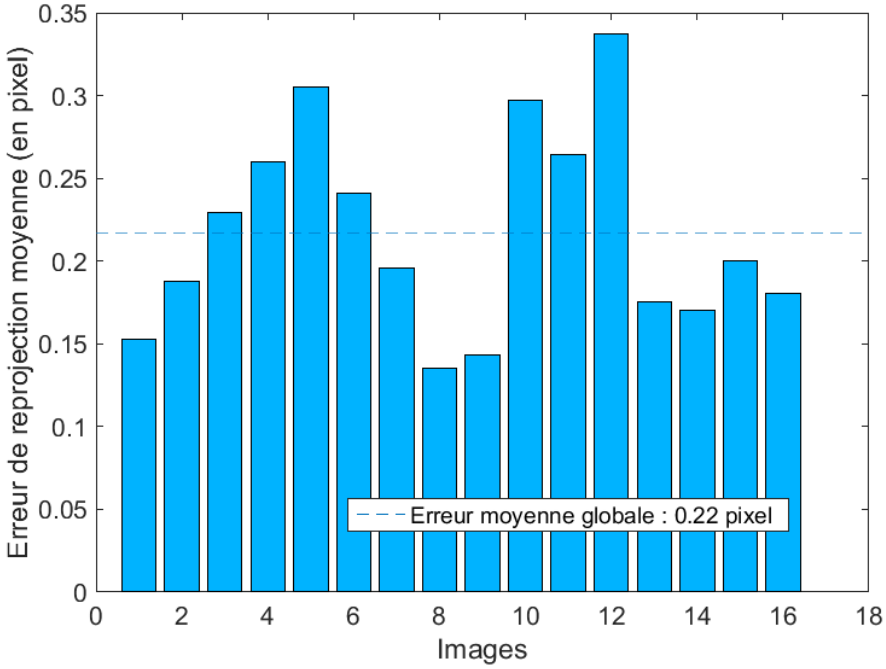


Figure 79. Erreur de reprojection des amers pour le calibrage grâce à la mire à damier par transmission.

3.3.1.2. Mire de type grille

Une grille en plastique a également été testée pour réaliser une mire. Elle est composée de 11×12 carrés dont les centres sont espacés de 7 mm (cf. Figure 80). La grille est positionnée devant le corps noir (cf. Figure 81). Les barreaux de la grille bloquent le rayonnement thermique du corps noir.

Pour détecter les amers de cette mire, la fonction *detectCheckerBoardPoints* ne peut pas être utilisée car les amers sont différents de ceux d'une mire à damier. Nous avons développé notre propre méthode. Il a été choisi de détecter les centres des carrés plutôt que les coins car les bords ne sont pas suffisamment nets sur certaines images. Les barycentres des carrés sont détectés automatiquement par seuillage.

Puis une interface permet de classer à la main les amers détectés selon la convention utilisée par Matlab. Dans cette convention, la numérotation des amers commence en haut à gauche du rectangle composé des $n \times m$ (avec $n < m$) amers (cf. Figure 82). La numération est incrémentée de haut en bas et de gauche à droite.

La focale de la caméra est évaluée à 713 pixels. Sachant que la taille du pixel vaut 17 µm, cela équivaut à une focale de 12.12 mm. La position estimée du centre optique est $[u, v]^T = [224{,}332]^T$. Pour rappel, la caméra possède une matrice de pixel au format 640×480 $pixels$ et une distance focale optique (c'est-à-dire la véritable distance focale) de 12 mm. Les coefficients de distorsion sont $k_1 = -0.40$ et $k_2 = 0.24$. L'erreur de reprojection associée à ce calibrage est illustrée par Figure 84. En moyenne elle vaut 0.17.

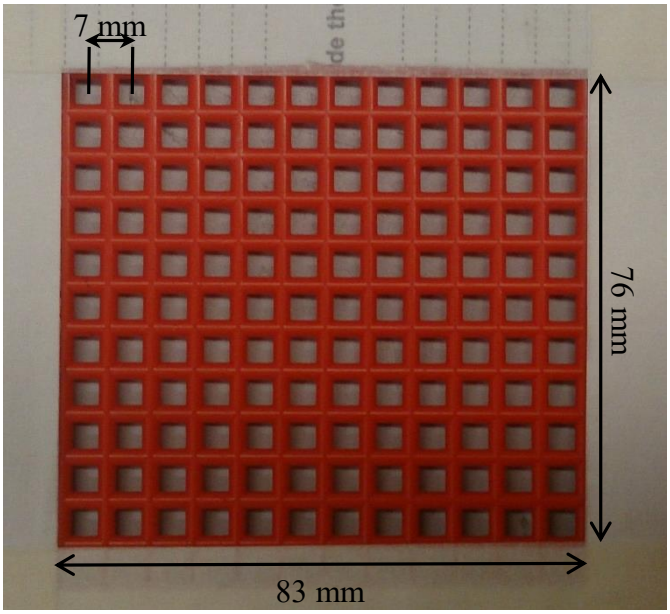


Figure 80. Grille en plastique utilisée pour le calibrage géométrique.

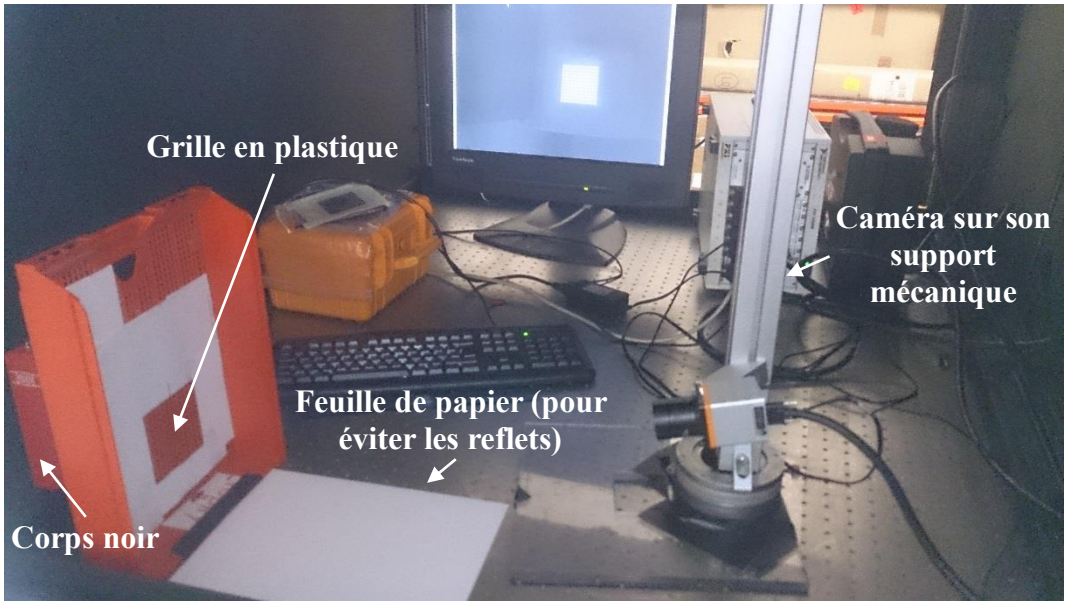


Figure 81. Disposition du matériel pour le calibrage géométrique gâce à une mire type grille.

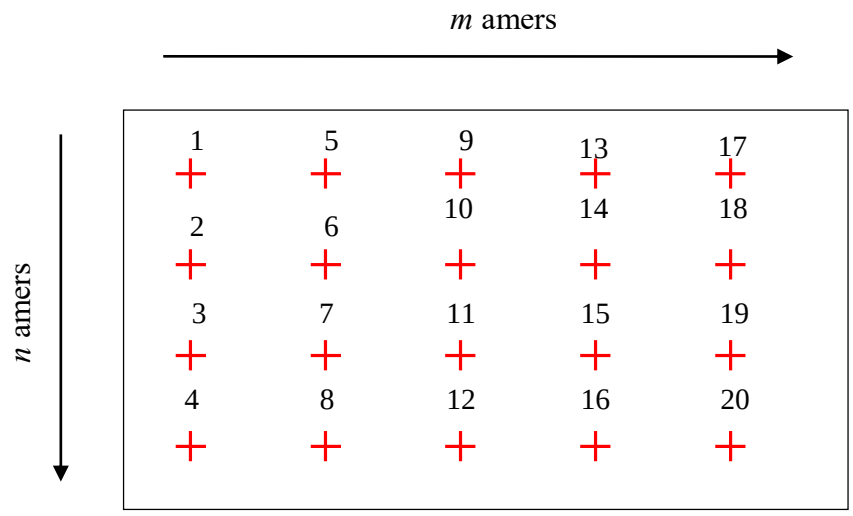


Figure 82. Convention permettant d'associer les coordonnées des amers détectés sur les images de la mire avec les coordonnées monde de la mire.

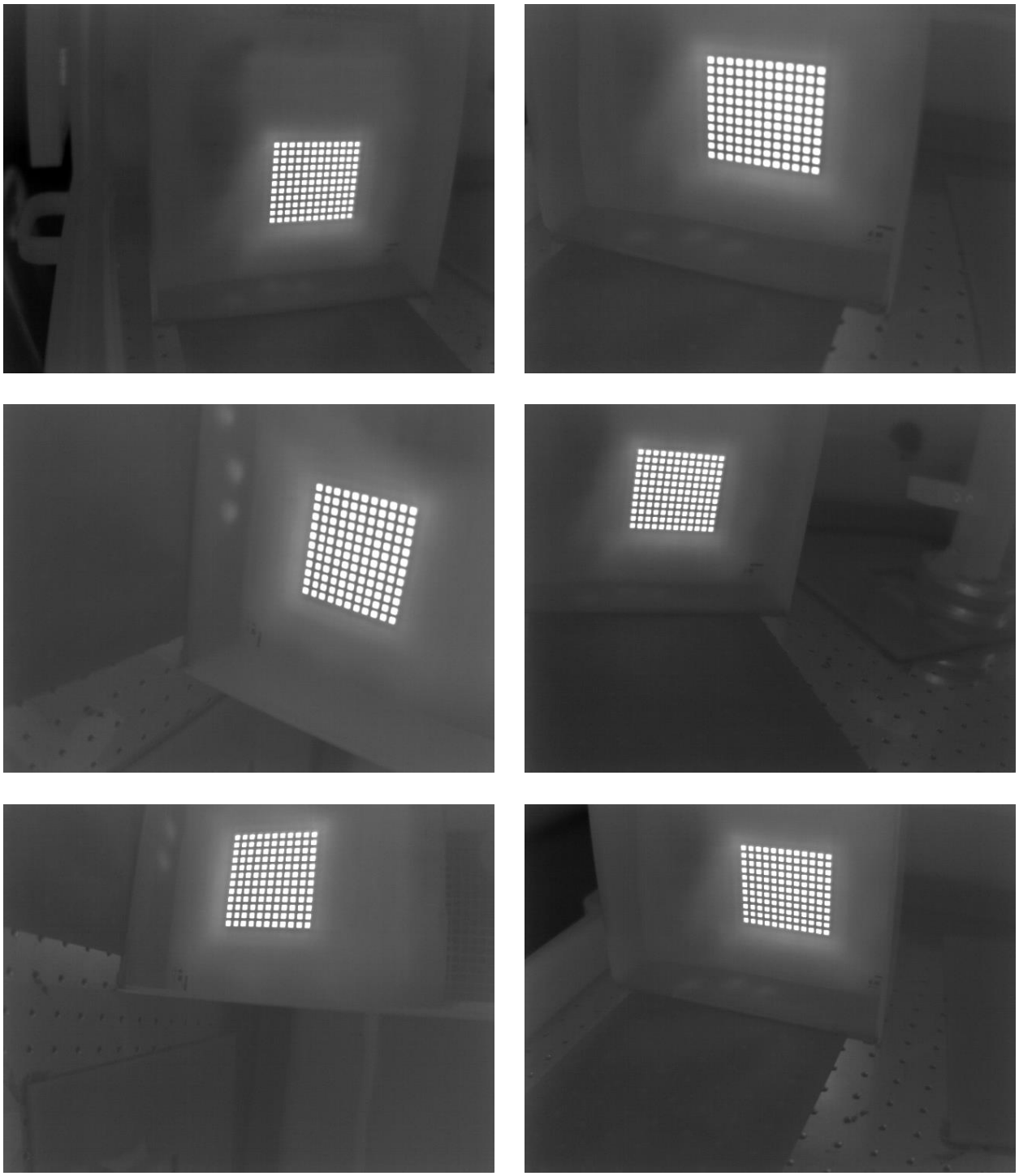


Figure 83. Six des vingt-et-une images d'une grille en plastique devant un corps noir, acquises de différents points de vue, pour le calibrage géométrique.

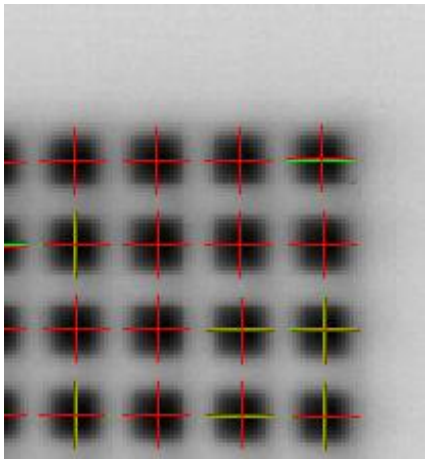


Figure 85. Zoom au niveau du coin supérieur droit de la première image de la Figure 83. Les croix vertes sont les centres des carrés noirs détectés. Les croix rouges sont les reprojections à l'aide des paramètres intrinsèques estimés par l'algorithme.

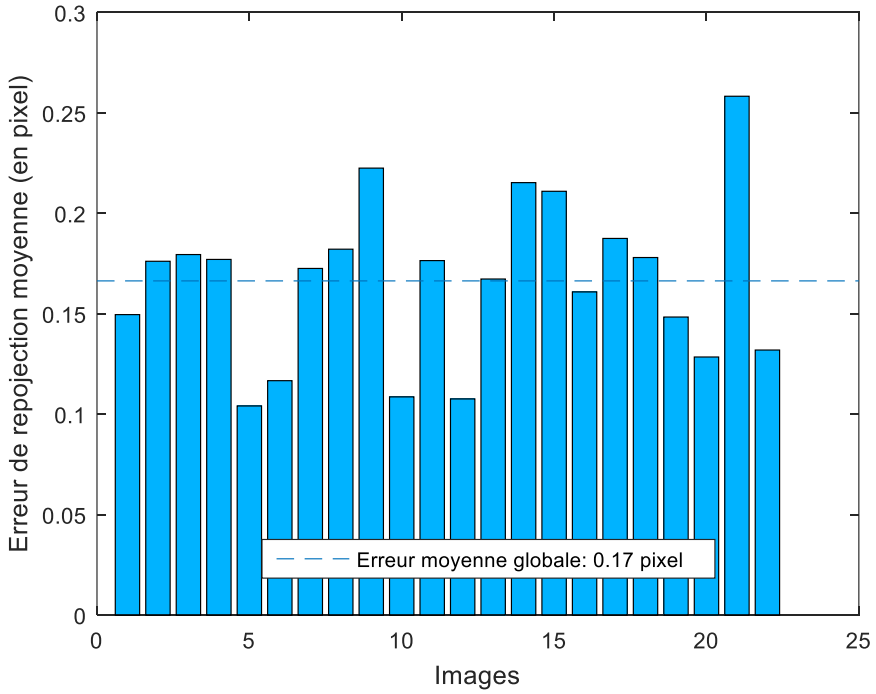


Figure 84. Histogramme des erreurs de reprojection moyenne par image. La moyenne globale sur les cinq images est représentée par le trait en pointillé.

3.3.1.1. Choix de la méthode de calibrage (mire ou grille)

La matrice de calibration obtenue grâce à la mire type grille sera préférée à la calibration obtenue grâce à la mire à damier car l'estimation de la focale semble plus cohérente avec distance focale de l'optique donnée par le fournisseur. $\boldsymbol{K}_{Gobi}$ désignera la matrice de paramètre intrinsèque dans la suite du manuscrit :

$$\boldsymbol{K}_{Gobi} = \begin{bmatrix} \alpha_0 & 0 & u_0 & 0 \\ 0 & \alpha_0 & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 713 & 0 & 332 & 0 \\ 0 & 713 & 224 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{3.20}$$

3.4. Estimation de la pose entre deux images successives

Lorsque que l'on estime la 'pose différentielle', c'est-à-dire entre deux images successives, on parle de méthode de **suivi de la pose**. Il est possible d'estimer la pose absolue par rapport à un état initial en sommant, au cours du temps, les poses différentielles. Nous avons implémenté une telle méthode. Les avantages et les inconvénients vont être mis en avant dans cette section. Nous conclurons sur la viabilité d'une telle approche pour une application de surveillance du conducteur.

La géométrie épipolaire décrit la relation géométrique entre deux images prises de deux points de vue différents, ou ce qui est équivalent, entre deux poses d'un objet pour une caméra fixe. Soit deux caméras notées A et B. $\boldsymbol{q}^A$ et $\boldsymbol{q}^B$ sont les projections d'un même point 3D Q. La contrainte épipolaire lie ces deux projections :

$$(\boldsymbol{q}^B)^T \boldsymbol{F}^{AB} \boldsymbol{q}^A = 0 \tag{3.21}$$

La matrice $\boldsymbol{F}$, appelée matrice fondamentale, dépend de la pose relative entre les deux caméras ainsi que des paramètres intrinsèques. Plusieurs méthodes permettent d'estimer $\boldsymbol{F}$ à partir de couples de points mis en correspondances [107,108]. Elles diffèrent les unes des autres essentiellement par le nombre de correspondances qu'elles utilisent (entre 5 et 8). Les différentes méthodes utilisées pour détecter et mettre en correspondance des points seront abordées plus en détails dans le chapitre 5. Les paramètres intrinsèques permettent d'estimer la matrice essentielle notée $\boldsymbol{E}^{AB}$. Cette dernière matrice contient l'orientation et la position relative entre les caméras A et B.

$$\boldsymbol{E}^{AB} = (\boldsymbol{K}_B)^T \boldsymbol{F}^{AB} \boldsymbol{K}_A \tag{3.22}$$

Dans notre cas, la caméra est unique et fixe donc $\boldsymbol{K}_B = \boldsymbol{K}_A = \boldsymbol{K}_{Gobi}$. Entre deux images successives, la position et l'orientation de l'objet ont évoluées. La matrice $\boldsymbol{E}^{AB}$ contient donc la pose relative du visage entre deux images successives.

Ce type d'approche nous est apparu, dans un premier temps, très intéressant car si on considère que le visage est déjà segmenté, aucune autre connaissance *a priori* sur la personne n'est nécessaire. En effet, en imagerie thermique, dans certaines conditions thermiques de l'habitacle, il est aisé de segmenter le visage. Il suffit qu'un contraste thermique suffisant entre le fond et le conducteur existe. Ces conditions sont très souvent vérifiées car la majorité des véhicules sont équipés d'une climatisation régulée en température. Non seulement, un contraste existe, mais en plus, le visage est plus chaud que le fond.

Remarque : en début de séquence, lorsque le conducteur entre dans le véhicule, il est possible d'utiliser la soustraction du fond pour la détection du visage.

Un problème récurrent en vision par ordinateur est la présence d'erreurs d'appariement. On parle d'*outliers*. Lorsque l'on met en correspondance des points entre deux images thermiques successives à une fréquence d'acquisition de l'ordre de 20Hz, il est tout à fait envisageable de considérer que le niveau de gris des correspondances correctes reste inchangé. Un avantage supplémentaire d'une approche de **suivi de la pose** est donc la possibilité de rejeter les *outliers* si les niveaux de gris varient trop d'une image à l'autre.

Il y a cependant un inconvénient majeur à ce type de méthode. Les erreurs d'estimation de la pose s'ajoutent au cours du temps. Cette raison est suffisante pour dire qu'il est inenvisageable d'estimer la pose du visage du conducteur à partir, uniquement, d'une méthode différentielle.

Nous recherchons donc une méthode capable d'identifier la pose avec précision à partir d'une seule image. On parle dans ce cas de méthode de **détection de la pose**. Dans un système final il est envisageable de combiner la rapidité d'une méthode de **suivi de la pose** avec la robustesse d'une méthode de **détection de la pose**.

Dans les chapitre 4 et 5, nous nous intéressons aux méthodes de **détection de la pose**. Comme cela a déjà été abordé dans le premier chapitre, les algorithmes de tracking du visage publiés dans la littérature scientifique utilisent souvent un maillage 3D du visage. Ce modèle 3D permettra de créer un avatar texturé qui, lui-même, permettra de créer une base d'images d'un ensemble de poses. L'image acquise par la caméra pourra donc être comparée exhaustivement à toutes les images de la base pour en déduire une estimation de la pose absolue.

3.5. Maillages 3D texturés du visage et création d'une base d'images de synthèse

La modélisation en 3D du visage implique le choix d'un maillage 3D et d'une texture. L'objectif est de rendre le modèle le plus fidèle à la réalité pour faciliter les algorithmes d'estimation de la pose. Nous allons commencer par présenter une manière couramment utilisée pour modéliser un objet en 3D. Nous montrerons également les modèles choisis et nous détaillerons le choix de la position du repère attaché au visage. Nous expliquerons ensuite la phase de plaquage de texture. Puis nous terminerons en précisant les paramètres des projections perspectives utilisées pour la création de la base d'images de synthèse.

3.5.1. La modélisation en 3D du visage grâce à un maillage

3.1. Définition des *faces* et des *vertices*

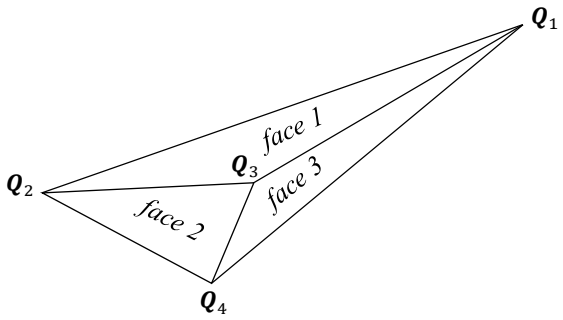


Figure 86. Représentation de trois *faces* et des *vertices* qui les définissent.

Nous considérons que le visage est un objet rigide. Les modèles utilisés le seront également. Un modèle 3D est généralement composé de *faces* et de sommets que l'on appellera par leur dénomination anglophone par la suite : les *vertices*. Les *faces* sont définies par trois *vertices* (cf. Figure 86).

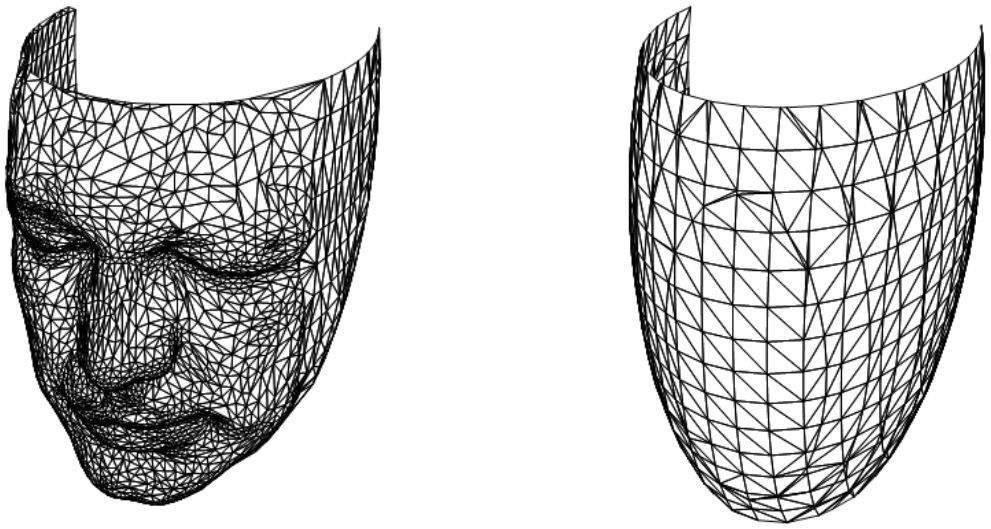
Un modèle 3D est souvent défini par deux tableaux. Le premier tableau répertorie les coordonnées 3D des *vertices*. Une ligne décrit un *vertex*, les colonnes contiennent les coordonnées X,Y et Z. Le second tableau défini les *faces*. Sur une ligne de ce dernier tableau, on retrouve trois indices qui permettent d'identifier trois *vertices* mis en jeux dans une *face*. Ces indices correspondent aux lignes du premier tableau (cf. Figure 87). L'indice k sera utilisé pour reconnaitre un *vertex* parmi les N qui définissent un modèle. $\boldsymbol{Q}_k = [X_k, Y_k, Z_k]^T$ définit complétement un *vertex*. L'ordre des *vertices* dans le tableau des *faces* indique le sens de la normale à celle-ci. Dans la Figure 86, les normales aux trois *faces* ont le même sens.

	Vertices		
$\boldsymbol{Q}_1$	X_1	Y_1	Z_1
$\boldsymbol{Q}_2$	X_2	Y_2	Z_2
$\boldsymbol{Q}_3$	X_3	Y_3	Z_3
$\boldsymbol{Q}_4$	X_4	Y_4	Z_4
	…		
$\boldsymbol{Q}_k$	X_k	Y_k	Z_k
	…		
$\boldsymbol{Q}_N$	X_N	Y_N	Z_N

	faces		
face 1	2	1	3
face 2	2	3	4
face 3	1	4	3
	…		
face M	…	…	…

Figure 87. Les tableaux de *vertices* et de faces définissent un modèle 3D du visage. Comme illustré sur la Figure 86, la *face 1* est définie par $\boldsymbol{Q}_2, \boldsymbol{Q}_1$et $\boldsymbol{Q}_3$.

3.5.1.2. Un maillage personnalisé et un maillage générique



a) b)

Figure 88. Maillages 3D du visage, a) maillage personnalisé obtenu grâce au scanner 3D de la société Artec3D, b) maillage ellipsoidal.

Deux maillages 3D ont été créés au cours de ces travaux. Le premier est adapté à la géométrie 3D de mon visage, le second modèle, grossier, peut être utilisé pour n'importe quel conducteur.

Pour créer le modèle adapté, nous avons utilisé un scanner 3D *Spider* de la société *Artec3D* (cf. Figure 88 a). Il projette un motif lumineux structuré (dans le proche infrarouge) sur l'objet à scanner et réalise des images de plusieurs points de vue. Le positionnement 3D des *vertices* $\boldsymbol{Q}_k$ est précis jusqu'à 50 µm lorsque le temps de préchauffe est respecté (30 minutes pour le modèle que nous avons utilisé, 3 minutes pour les nouvelles versions). De plus, ces performances sont atteintes lorsque la distance entre le scanner et l'objet est compris entre 17 et 35 cm. Le champ de vision horizontal vaut 30° et celui vertical vaut 21°. A 35 cm, le champ de réception est un rectangle de 18 × 14 cm. En pratique, pour acquérir la 3D du visage, la petite dimension du champ de réception nécessite de *scanner* lentement. Finalement ce produit est très précis mais nécessite de prendre des précautions lors du *scan*. Il est également très onéreux. Le produit *Eva* de la même société semblerait plus adapté à notre application puisque le champ de réception vaut 53.6 × 37.1 cm pour une distance entre l'objet et le scanner de 1 m. La précision sur le positionnement des *vertices* est alors de 100 µm. Le produit *Eva* est également très onéreux.

D'autres produits, moins coûteux et moins précis existent. Par exemple, la société *3DSystems* propose le produit appelé *Sense 2* [109]. Le champ de réception de *Sense 2* atteint 1.33 × 1.75 m pour une distance entre l'objet et le scanner de 1.6 m. Sa précision sur le positionnement des *vertices* est proche de 1 mm. Avec des spécifications techniques du même ordre, on trouve également la gamme de produits *Carmine* de la société *Primesense* (la société *Primesense* a été acquise par la société *Apple* en 2013) [110]. Encore plus économique, des travaux proposent une modélisation 3D du visage grâce à une Kinect [111]. Enfin, avec une simple caméra RGB, on trouve dans la littérature scientifique des méthodes qui proposent d'adapter

automatiquement un modèle 3D générique aux spécificités d'un individu en utilisant plusieurs vues [112]. Il est également possible d'adapter manuellement un modèle moyen paramétrable. Le modèle *Candide-3* téléchargeable depuis la référence [113] est très populaire (cf. Figure 89). Ce modèle est utilisé pour estimer la pose d'un visage dans la référence [114] et dans le logiciel commercial *Visage Technologies* [115]. Les paramètres sont, par exemples, l'échelle globale, la hauteur et l'écartement des yeux, la position l'épaisseur et la largeur de la bouche…

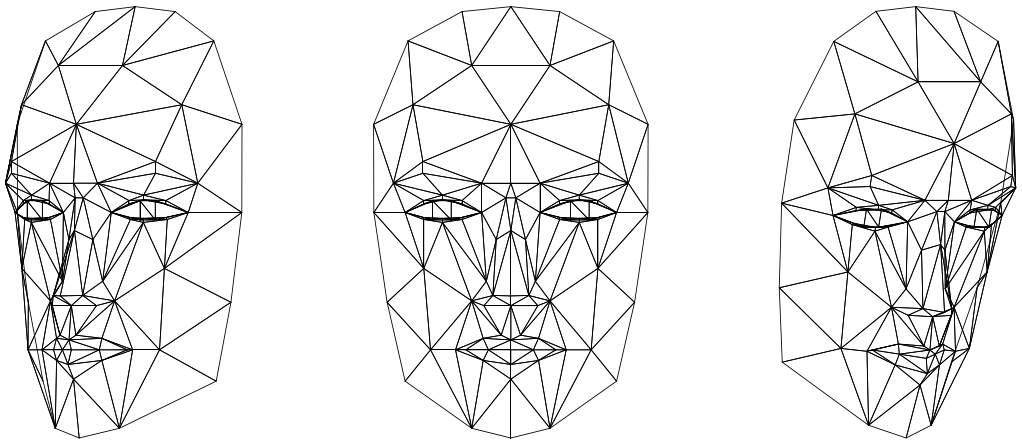


Figure 89. Modèle paramétrable *Candide-3* disponible depuis la référence [113].

La phase de personnalisation de la géométrie 3D peut être contraignante pour certaines applications. Des modèles plus basiques peuvent alors être utilisés. Dans la référence [116] un simple plan modélise le visage. L'efficacité de ce type de modèle est limitée en cas de rotation importante hors du plan. Dans la référence [117] un modèle cylindrique est utilisé pour gérer des angles plus importants. Dans la référence [118] un modèle ellipsoïdal est plutôt utilisé. Nous avons souhaité tester ce dernier type de modèle car il nous semble plus proche de la géométrie réelle d'un visage par rapport à un modèle cylindrique, sans pour autant ajouter de la complexité dans l'implémentation (cf. Figure 88 b). Les paramètres de l'ellipsoïde ont été évalués par ajustement avec le modèle 3D précis obtenu avec le scanner *Artec3D*.

Pour résumer, la Figure 90 illustre un classement des modèles 3D selon leurs précisions géométriques. Nous distinguons les modèles personnalisés et les modèles génériques. Nous avons choisis de tester un modèle dans chacune de ces catégories.

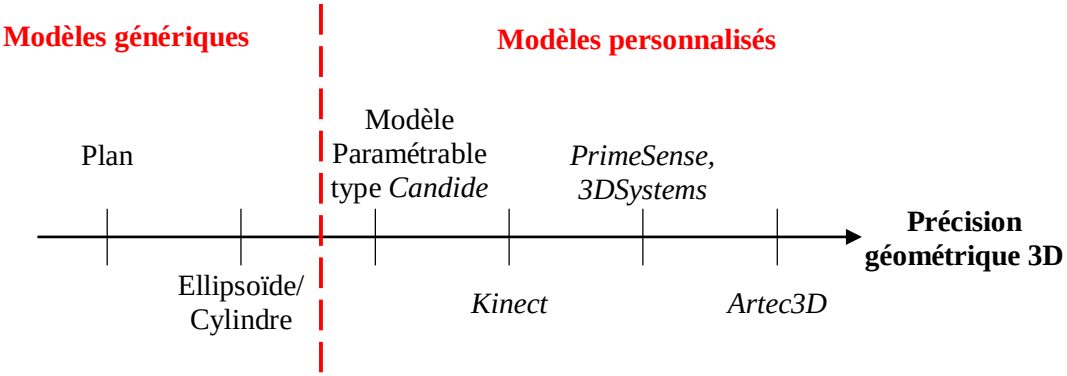


Figure 90. Modèles 3D classés selon leur précision géométrique 3D.

3.5.1.3. Une caméra virtuelle et un repère attaché à l'objet

La manipulation d'un modèle 3D implique (i) l'utilisation d'une *caméra virtuelle* et (ii) un repère attaché au modèle. Abordons d'abord la *caméra virtuelle*. Que cela soit **1.** pour l'étape de plaquage de texture sur le maillage 3D (cf. section 3.5.2) ou **2.** pour la création d'images de synthèse (cf. section 3.5.3), il est nécessaire de simuler l'image du maillage 3D, c'est-à-dire de projeter les *vertices* sur le plan image. Le terme *caméra virtuelle* désigne donc une projection perspective donnée.

1. Lors de l'étape de plaquage de texture (cf. section 3.5.2), on va extraire le niveau de gris de l'image issue de la caméra thermique *Gobi 640 CL*. Pour cela nous devrons projeter le maillage 3D sur l'image réelle de la caméra thermique. Les paramètres intrinsèques de la caméra seront donc utilisés pour cette projection.

2. Dans l'étape de la création de la base d'images de synthèse (cf. section 3.5.3), on souhaite obtenir des images proches de celles obtenues par la caméra *Gobi 640 CL*. Le maillage 3D sera également projeté dans le plan image grâce aux paramètres intrinsèques de la caméra.

Pour conclure, la *caméra virtuelle* est définie par son repère noté $\boldsymbol{O^V X^V Y^V Z^V}$ et ses paramètres intrinsèques $\boldsymbol{K^V = K_{Gobi}}$. Pour le champ de la caméra, nous utilisons la donnée du fournisseur. Le champ de vue horizontal $HFOV$ (de l'anglais *horizontal field of view*) spécifié par *Xenics* est $HFOV = 53°$. Nous définirons les champs horizontaux et verticaux de la caméra virtuelle à 53°.

Nous notons le repère orthonormé, le repère attaché au modèle 3D, $\boldsymbol{O^O X^O Y^O Z^O}$. Les angles *yaw* (lacet), *pitch* (tangage), et *roll* (roulis) désignent les angles autour des axes respectifs $\boldsymbol{Y^O}$, $\boldsymbol{X^O}$ et $\boldsymbol{Z^O}$ (cf. Figure 91). Concernant la position des axes de rotation, l'approximation suivante a été utilisée dans la littérature [119] : la tête est modélisée par une sphère qui peut tourner autour d'un point qui se situe au centre de cette sphère. Le vecteur $\boldsymbol{X^O}$ est colinéaire à l'axe des yeux. Le vecteur $\boldsymbol{Y^O}$ passe par le centre de la bouche, le centre du

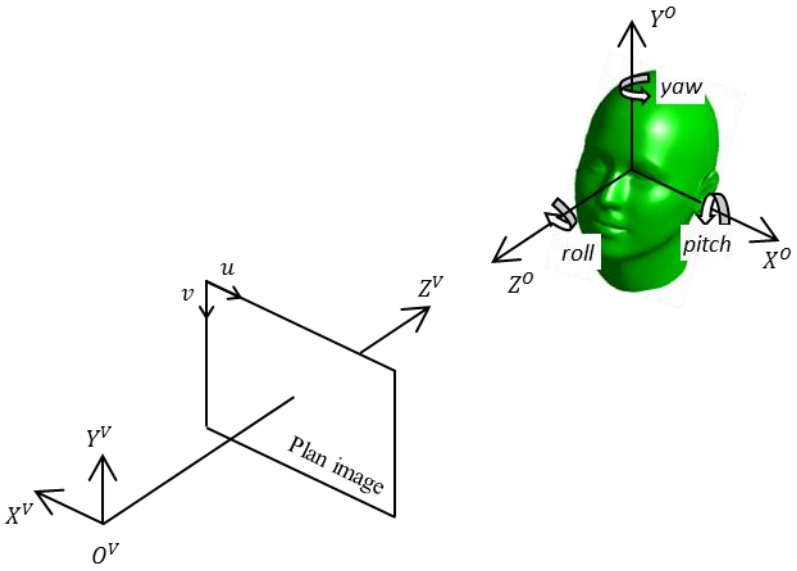


Figure 91. Notations utilisées pour la *camera virtuelle* et le repère attaché à l'objet.

nez et entre les deux yeux. Enfin le vecteur $\boldsymbol{Z}^O$ est colinéaire à l'axe qui part de la base du nez jusqu'à sa pointe (cf. Figure 91).

L'erreur de position du repère objet :

La position du centre du repère $\boldsymbol{O}^O$ est dictée par des aspects physiologiques humains complexes. En réalité, les axes de rotation peuvent se décaler en fonction de la nature du mouvement.

Comme nous allons le décrire dans la section 3.5.3, le modèle 3D texturé va subir une projection perspective afin de synthétiser une image qui doit être fidèle à la réalité. La position de $\boldsymbol{O}^O$ est importante dans le cas où l'on souhaite appliquer une rotation au modèle 3D suivi d'une projection perspective. En effet, elle impacte la taille des segments de l'objet 3D projeté sur le plan image (cf. Figure 92). On peut remarquer sur la Figure 92 que dans la situation b) l'objet est décalé de $e = d.\tan(\theta)$. Ainsi la taille de la projection perspective du vecteur $\overrightarrow{\boldsymbol{Q}_1\boldsymbol{Q}_2}$ dépend de la position du centre du repère objet. Pour illustrer cet aspect prenons un exemple et faisons l'application numérique. Intéressons-nous à la dimension d'un œil sur le modèle personnalisé. Soit $\boldsymbol{Q}_1$ et $\boldsymbol{Q}_2$ deux points situés aux extrémités horizontales. Nous notons $\overline{\boldsymbol{q}_1\boldsymbol{q}_2}$ la taille de la projection du vecteur $\overrightarrow{\boldsymbol{Q}_1\boldsymbol{Q}_2}$. Nous allons appliquer une rotation de *yaw* = 35°.

- Dans un premier cas le centre de rotation est situé au milieu du modèle 3D. On obtient $\overline{\boldsymbol{q}_1\boldsymbol{q}_2} = 76\ pixels$.
- Dans un second cas, le centre de rotation est situé 10 cm en arrière du centre du modèle (en direction de la nuque). On obtient $\overline{\boldsymbol{q}_1\boldsymbol{q}_2} = 72\ pixels$.

La dimension d'un œil peut varier de 4 $pixels$ en fonction du choix du centre de rotation de la tête. Voici la conséquence d'une telle erreur : lorsque le centre de rotation est situé au milieu du modèle, la dimension horizontale de l'œil $\overline{\boldsymbol{q}_1\boldsymbol{q}_2}$ peut atteindre 72 $pixels$ si le visage est orienté d'un angle *yaw* qui vaut 45°. Ainsi, un visage orienté d'un angle $yaw = 35°$ avec un centre de rotation placé 10 cm en arrière du centre

du modèle ressemble, sous certains aspects (dans ce cas il s'agit de la taille horizontale de l'œil), à un visage orienté d'un angle $yaw = 45°$ avec un centre de rotation au centre du modèle.

La position des axes de rotation de la tête n'est pas connue avec précision car un grand nombre de vertèbres sont impliquées dans la liaison entre le coup et la tête. Elles provoquent une redondance des possibilités de rotation. De plus, il existe des variabilités inter-individu. Les réflexes vestibulo-oculaires (*VOR* pour *vestibulo-ocular reflex* dans la littérature anglophone) modulent la position des axes de rotation afin de stabiliser l'image sur la rétine de l'œil [119,120]. Il y a donc une différence entre les mouvements volontaires de la tête et les oscillations involontaires (ou reflexe). Ainsi la position des axes de rotation peut être modulée.

Les mouvements à faible amplitude (*pitch* <15° et *yaw* <25°) et à haute fréquence (entre 1 et 2 Hz) ont été spécifiquement étudiés dans la référence [121]. L'axe $\boldsymbol{X}^{O}$ (angle *pitch* sur la Figure 91) se situe en dessous de l'axe interaural (l'axe entre les deux oreilles, cf. Figure 93). Lorsque la fréquence augmente, il a tendance à se rapprocher de cet axe. De même, en posture assise (par rapport à la posture debout), il se rapproche également de l'axe interaural. En moyenne, la position de l'axe $\boldsymbol{X}^{O}$ peut varier de 26 mm en hauteur et de moins de 3 mm en profondeur.

Toujours dans la référence [121], la position de l'axe $\boldsymbol{Y}^O$ (angle *yaw* sur la Figure 91) est estimée pour des mouvements répétés à une fréquence de 1 Hz. Elle est indépendante de la posture (assis ou debout) et de l'amplitude du mouvement (des mouvements de $\pm 10°$ et $\pm 20°$ ont été testés). L'axe de rotation est situé environ 10 mm derrière l'axe interaural.

Dans la référence [120] l'amplitude des rotations selon l'axe $\boldsymbol{Y}^O$ (angle *yaw*) va jusqu'à 40°. Cette étude confirme l'invariance de la position l'axe $\boldsymbol{Y}^O$ en fonction de l'amplitude de l'angle *yaw*. Elle confirme également la dépendance de la position de l'axe $\boldsymbol{X}^O$ (angle *pitch*) à l'amplitude de l'angle *pitch*. On constate notamment une dépendance forte sur les mouvements dit de flexion (le participant regarde vers le sol). L'axe de rotation peut ainsi se déplacer de près de 60 mm vers le bas. Ce cas d'usage est à prendre en compte car il peut correspondre à un conducteur qui regarde son *smartphone*.

Dans le chapitre 5, notre test inclus des rotations qui vont de -60° à 60° en *yaw* et de -20° à +20° en *pitch*. Envisageons un angle *pitch* = 20° et mesurons l'épaisseur de la projection de bouche dans un cas ou l'axe de rotation serait idéalement placé et dans un cas où il serait positionné avec une erreur de 60 mm. L'épaisseur de la bouche varie de 0.6 pixels. C'est donc assez faible.

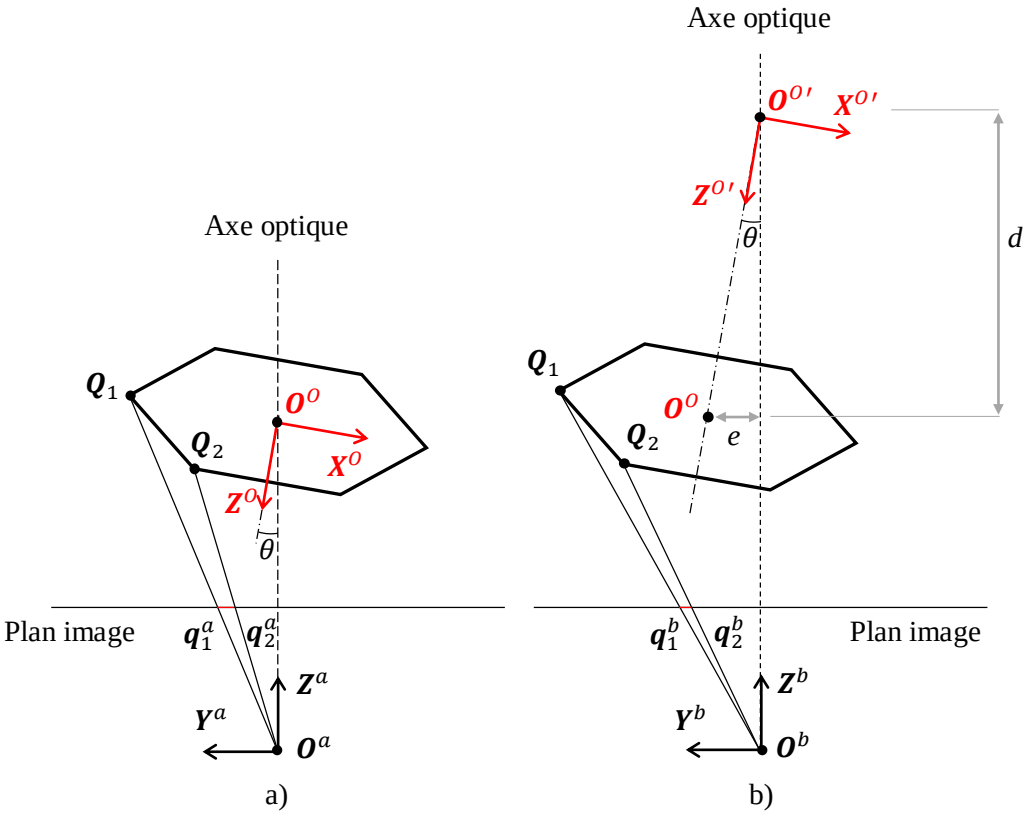
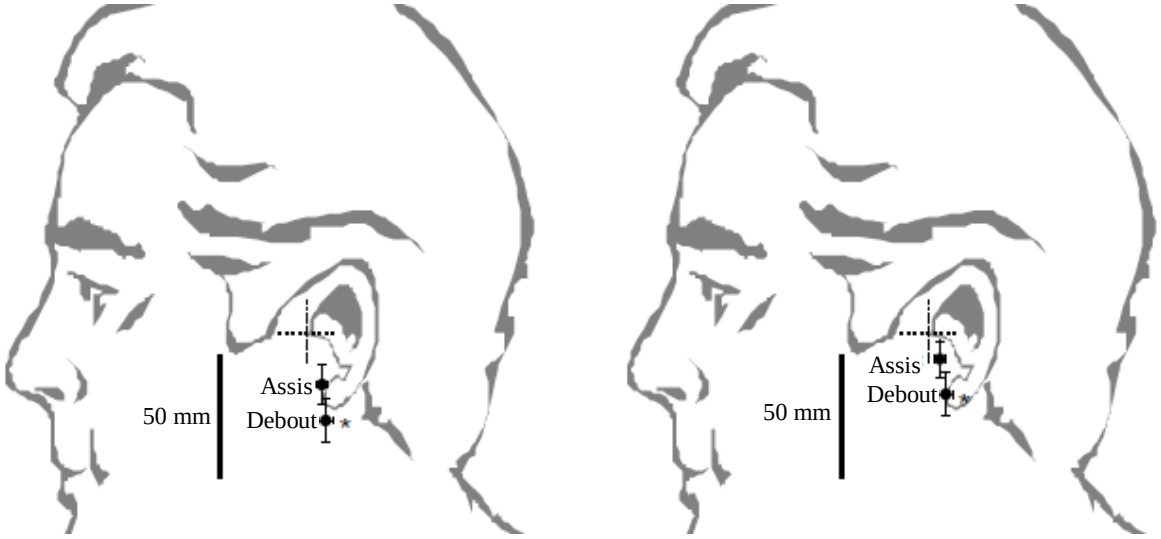


Figure 92. L'hexagone en trait légèrement épais est l'objet imagé. Une rotation d'angle θ lui est appliquée dans les deux situations. Dans la situation a) $\boldsymbol{O}^O$ est positionné au centre de l'objet. Dans la situation b) $\boldsymbol{O}^O$ est positionné derrière l'objet. Nous souhaitons illustrer l'impact du choix de la position du centre de rotation sur la projection perspective. On remarque en effet que $\overline{\boldsymbol{q}_1^a \boldsymbol{q}_2^a} > \overline{\boldsymbol{q}_1^b \boldsymbol{q}_2^b}$.

Dans la suite de la thèse, nous modéliserons la tête comme un objet rigide avec un unique centre de rotation, placé au milieu de l'axe interaural.



a) b)

Figure 93. Les points en noir représentent la position moyenne (calculée à partir d'un minimum de quatre participants) de l'axe de rotation $\boldsymbol{X}_0$ (angle *pitch*) [121]. Des mouvements oscillatoires purs de haut en bas à une fréquence de 1 Hz a), et de 2 Hz b) ont été effectués en posture assise ou debout. La barre d'erreur est l'écart type sur la position de l'axe (il est calculé à partir d'un minimum de quatre participants). La croix en pointillé symbolise l'axe interaural.

3.5.2. Extraction et plaquage de la texture sur le maillage 3D

L'objectif de cette section est de décrire la méthode permettant d'extraire la texture de l'image thermique pour la positionner sur le maillage 3D. Le choix a été pris d'utiliser la texture provenant de trois points de vue du visage : face, profil droit et profil gauche (cf. Figure 94). Ceci garantira un fonctionnement des algorithmes à des rotations importantes du visage selon l'angle *yaw* (jusqu'à ±60°).

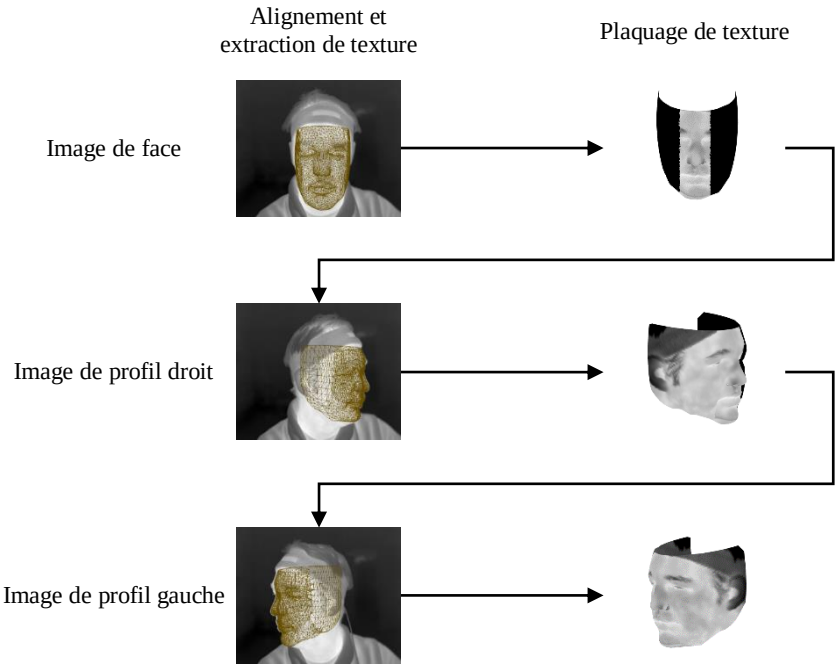


Figure 94. Procédure d'alignement et d'extraction de la texture de trois points de vue du visage (face, profil droit et profil gauche).

On définit trois transformations représentant les trois poses (la pose est l'orientation et la position d'un objet) du visage dans les trois images : face, profil doit et profil gauche. Elles sont respectivement notées :

- $\left[\boldsymbol{R}^{face}\middle|\boldsymbol{t}^{face}\right]$
- $\left[\boldsymbol{R}^{droit}\middle|\boldsymbol{t}^{droit}\right]$
- $[\boldsymbol{R}^{gauche}|\boldsymbol{t}^{gauche}]$

Le maillage 3D après application de ces transformations est illustré sur la Figure 95.

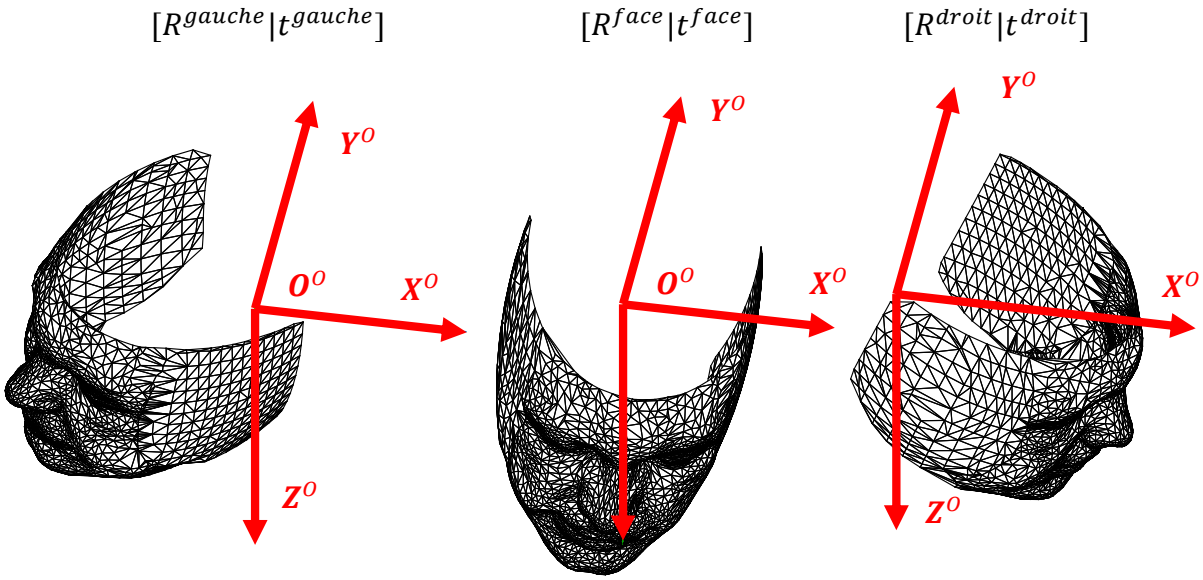


Figure 95. Alignement du maillage 3D avec les trois images utilisées pour l'extraction de la texture, a) le maillage subit la rotation et la translation $[\boldsymbol{R}^{gauche}|\boldsymbol{t}^{gauche}]$, b) le maillage subit la transformation $[\boldsymbol{R}^{face}|\boldsymbol{t}^{face}]$, et c) le maillage subit la transformation $[\boldsymbol{R}^{droit}|\boldsymbol{t}^{droit}]$.

Pour extraire la texture de l'image de face, les *vertices* sont projetés sur l'image acquise grâce à la projection perspective $\boldsymbol{K}_{Gobi}[\boldsymbol{R}^{face}|\boldsymbol{t}^{face}]$. Les six degrés de liberté contenus dans la matrice $[\boldsymbol{R}^{face}|\boldsymbol{t}^{face}]$ sont établis semi-manuellement. Nous commençons d'abord par segmenter grossièrement le visage. Pour faire cela, une interface a été créée permettant de relever des points au bord du visage manuellement. Ensuite l'ellipse qui ajuste au mieux ces points est estimée (cf. Figure 96). Le demi grand axe de l'ellipse est noté a et le demi petit axe est noté b. Les coordonnées du centre de l'ellipse (en pixels) sont notées x_e, y_e. Ces paramètres vont permettre de faire une première estimation des composantes de $\boldsymbol{t}^{face}$.

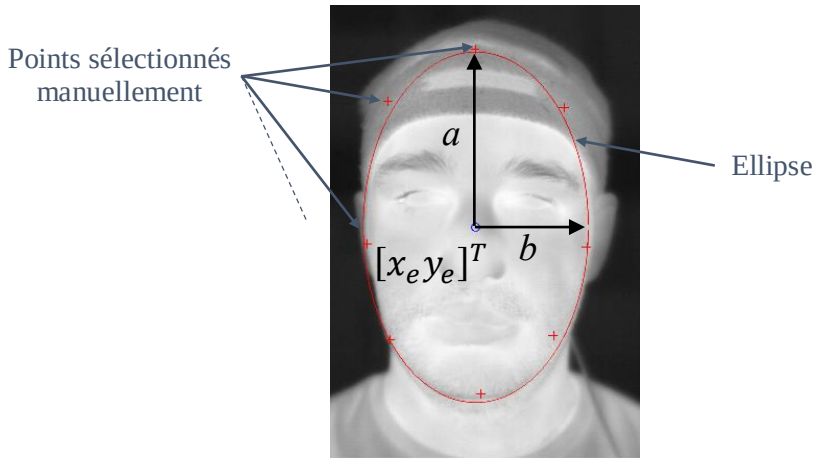


Figure 96. Segmentation semi-automatique du visage pour le point de vue de face.

L'estimation de la demi largeur du visage, en valeur absolue, sur l'image est représentée par le demi petit axe *b*. Elle permet d'établir la distance entre le repère caméra et le repère objet (c'est-à-dire le repère attaché au visage) t_z^{face} grâce à la relation appelée grandissement transversal :

$$|b| = \left| \alpha_0 \frac{max(X_k) - min(X_k)}{2.\, t_z^{face}} \right| \tag{3.23}$$

Les coordonnées X_k sont les *vertices* du maillage 3D. La largeur du visage dans le repère objet en valeur absolue est $|\max(X_k) - \min(X_k)|$. La focale exprimée en nombre de pixels est le terme α_0 issue de la phase de calibrage géométrique (cf. équation (3.20)).

En faisant l'approximation que la projection du centre du repère objet $\boldsymbol{O}^O$ correspond au centre de l'ellipse de coordonnées $[x_e, y_e\,]^T$ et grâce aux équations (3.6) et (3.7), on peut déduire une première estimation de t_X^{face} et t_Y^{face} :

$$\begin{aligned} x_e &= \alpha_0 \frac{t_X^{face}}{t_z^{face}} \\ y_e &= \alpha_0 \frac{t_Y^{face}}{t_z^{face}} \end{aligned} \tag{3.24}$$

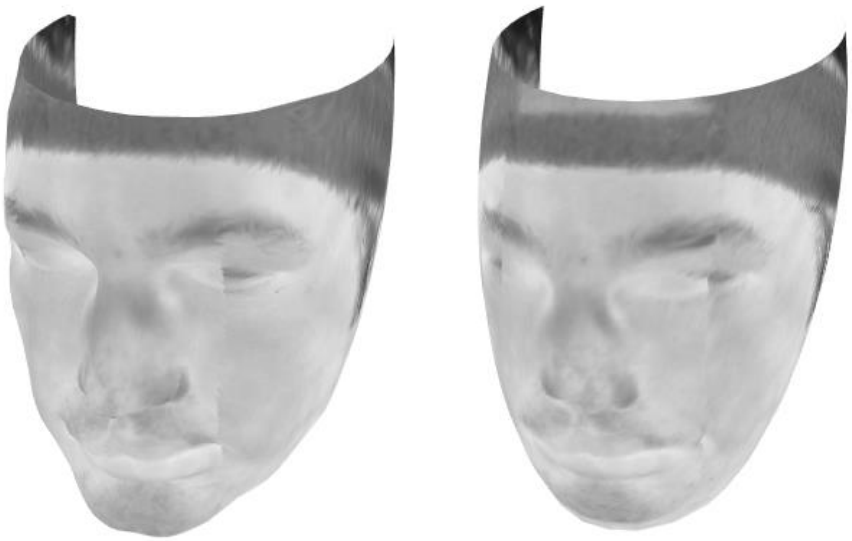
Il est ensuite possible d'affiner manuellement $\boldsymbol{t}^{face}$ et $\boldsymbol{R}^{face}$.

Une fois l'alignement du modèle 3D réalisé avec l'image de face, les valeurs des pixels de l'image (en *Adu* de l'anglais *analog digital output*) aux positions des projections des *vertices* ($\boldsymbol{q}_k$) vont être associées aux indices k. La couleur d'une *face* est le résultat d'une interpolation bilinéaire à partir des trois valeurs des *vertices* qui la définissent.

Pour extraire la texture d'une image de profil droit et gauche, on utilise les angles obtenus à partir d'une centrale inertielle portée par le participant (nous reviendrons sur les caractéristiques de cet appareil lorsque nous présenterons la procédure d'évaluation des algorithmes d'estimation de la pose). On obtient alors $\boldsymbol{R}^{droit}$ et $\boldsymbol{R}^{gauche}$. Il est souvent nécessaire de modifier manuellement $\boldsymbol{t}^{droit}$ et $\boldsymbol{t}^{gauche}$ car les mouvements du visage sont généralement la combinaison d'une rotation et d'une translation. Une fois l'alignement du modèle 3D réalisé avec l'image de profil droit et celle de profil gauche, on répète l'opération de plaquage de texture aux *vertices*. Le résultat obtenu est illustré sur la Figure 97.

3.5.3. Création de la base d'images de synthèse

Grâce au modèle 3D texturé, nous allons pouvoir synthétiser des images différemment orientées et étiquetées. Dans cette section, le repère de la caméra virtuel est fixe (par rapport au repère monde), et l'objet virtuel est en mouvement. Les coordonnées 3D des *vertices* $\boldsymbol{Q}_k^O$ sont définies initialement dans le repère objet. Pour décrire le mouvement de l'objet, nous allons utiliser un nouveau repère noté $\boldsymbol{O}^v\boldsymbol{X}^v\boldsymbol{Y}^v\boldsymbol{Z}^v$. Il s'agit du repère objet multiplié par une matrice de rotation $\boldsymbol{R}^v$. Ainsi dans ce repère, les *vertices* s'expriment comme suit :



a) b)

Figure 97. Les deux maillages 3D texturés à partir de trois images thermiques du visage. Le maillage précis texturé a) et le maillage ellipsoïdal texturé b).

$$\begin{bmatrix} X_k^v \\ Y_k^v \\ Z_k^v \end{bmatrix} = \boldsymbol{R}^v \begin{bmatrix} X_k^O \\ Y_k^O \\ Z_k^O \end{bmatrix} \tag{3.25}$$

La matrice de rotation $\boldsymbol{R}^v$ est définie comme suit :

$$\boldsymbol{R}^v = Rot(Z, \psi = 0) Rot(Y, \varphi) Rot(X, \theta)$$

$$\boldsymbol{R}^v(\theta, \varphi) = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} cos(\varphi) & 0 & sin(\varphi) \\ 0 & 1 & 0 \\ -sin(\varphi) & 0 & cos(\varphi) \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\theta) & -sin(\theta) \\ 0 & sin(\theta) & cos(\theta) \end{bmatrix} \tag{3.26}$$

Une image de synthèse est caractérisée uniquement par le couple d'angle $\varphi = yaw$, $\theta = pitch$. Pour ces travaux de thèse, nous avons synthétisé 225 images pour lesquelles *yaw* varie de -60° à +60° par pas de 5° et *pitch* varie de -20° à +20° par pas de 5° (cf. Figure 98).

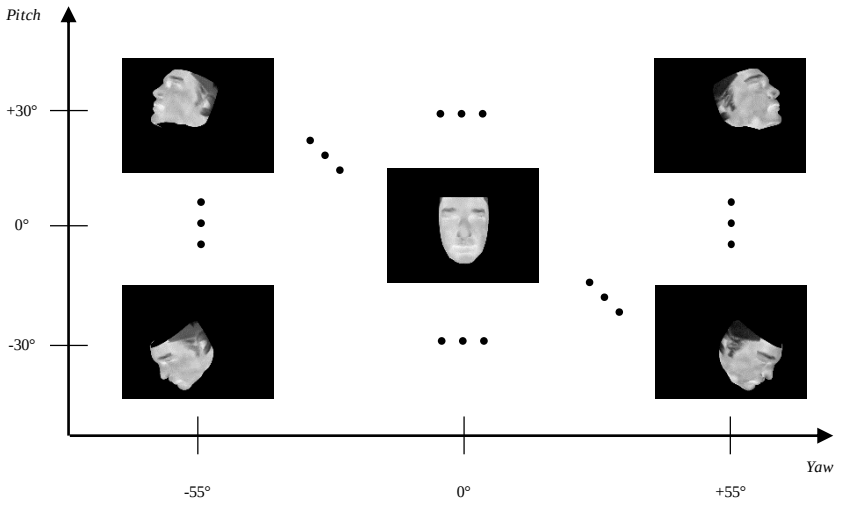


Figure 98. Base d'images de synthèse différemment orientées en (*yaw,pitch*)

Chapitre 4. Estimation de la pose basée sur une approche « problèmes inverses »

4.1. Introduction

Nous allons montrer dans ce chapitre comment exploiter la base d'images de synthèse thermiques d'un visage pour estimer deux angles qui définissent partiellement l'orientation du visage :

- l'angle *yaw* (lacet) par rapport au repère monde,
- l'angle *pitch* (tangage) par rapport au repère monde,

Remarque : on estime également la position [u,v] de la projection du visage même si nous ne l'utilisons pas dans la suite du chapitre.

- *la position selon l'axe horizontale par rapport à une position initiale,*
- *la position selon l'axe verticale par rapport à une position initiale.*

Nous allons utiliser un algorithme basé sur une approche « problèmes inverses ». Cette méthode peut être qualifiée de « globale » car tous les pixels du visage vont être utilisés contrairement à la méthode qui sera développée dans le Chapitre 5 qui est qualifiée de locale.

Il est tout à fait possible d'envisager d'estimer les six degrés de libertés qui composent la pose 3D avec des algorithmes « problèmes inverses » comme cela a été montré dans la référence [123]. Pour cela un algorithme plus complexe doit être implémenté. L'objectif de ce chapitre, plus modeste, est de montrer la faisabilité de ce type d'approche en imagerie thermique, tout en quantifiant la précision de l'estimation sur l'orientation.

Nous commencerons par présenter une méthode expérimentale d'estimation de l'orientation du visage du conducteur pour la construction de la « vérité terrain » (section 4.2). Puis nous détaillerons l'algorithme basé sur une approche « problèmes inverses » (section 4.3). Puis nous conclurons (section 4.4).

4.2. Construction de la « vérité terrain » pour l'estimation de l'orientation du visage

Nous avons mis en place une méthode pour évaluer l'orientation du visage du conducteur. Un appareil de mesure permet de fournir une « vérité terrain » sur les orientations. Il est également nécessaire de définir un scénario représentatif des cas d'utilisation d'un système d'estimation de l'orientation du visage dans une situation de conduite automobile.

4.2.1. La vérité terrain : une centrale inertielle

Afin de quantifier la précision de l'estimation des angles *yaw* (lacet) et *pitch* (tangage) nous avons utilisé une centrale inertielle comme cela a également été fait dans les références [124] et [125]. Nous avons utilisé la centrale inertielle *InertiaCube3* commercialisée par la société *InterSense* [126]. Celle-ci combine l'évaluation de composantes du champ gravitationnelle et du champ magnétique de la terre pour mesurer les trois angles. La Figure 99 présente le diagramme fonctionnel de la centrale inertielle.

La mesure des angles *roll* (roulis) et *pitch* (tangage) ne dérive pas car la force gravitationnel de la terre permet de réaliser une mesure absolue. Par contre, la mesure de l'angle *yaw* (lacet) ne peut pas se baser sur la force gravitationnelle (car un objet différemment orienté selon cet angle subit toujours la même force gravitationnelle). Cet angle est donc mesuré en intégrant dans le temps des mesures relatives à partir d'une remise à zéro manuelle. Cette méthode est sujette à une dérive. Les senseurs magnétiques permettent de corriger cette dérive en repérant le pôle nord magnétique. Ce procédé de compensation de la dérive peut être perturbé par l'environnement magnétique. Ainsi, nous n'avons pas réalisé de tests quantifiés à l'intérieur d'un habitacle automobile. Nous avons préféré utiliser un modèle simplifié du véhicule décrit à la section 4.2.2.

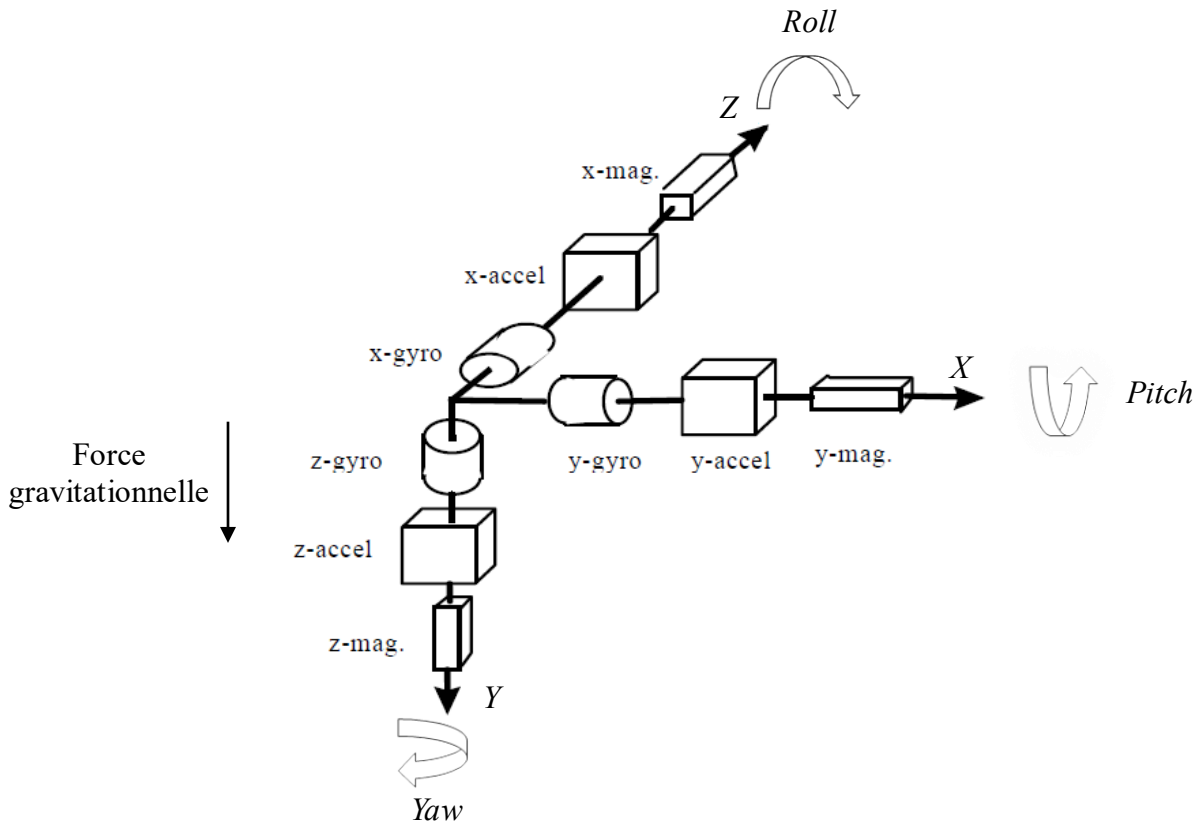


Figure 99. Diagramme fonctionnel de la centrale inertielle *InertiaCube3* de *InterSense* [126].

Lorsque l'environnement magnétique ne perturbe pas la mesure, le fournisseur annonce les précisions *RMS* suivantes :

- 1° en *yaw*,
- 0.25° en *pitch* et *roll*.

Pour comparer la mesure de la centrale inertielle à une estimation issue du traitement des images thermiques, il est nécessaire d'horodater les données. Nous avons développé un script grâce au logiciel *LabView* qui permet de lire, de copier et d'horodater les mesures de la centrale inertielle ainsi que les images de la caméra thermique *Gobi 640 CL*. Les angles de la centrale inertielle sont acquis et copiés à une fréquence d'environ 200 Hz et les images à une fréquence d'environ 15 Hz.

Pour associer un ensemble de trois données (*yaw*, *pitch*, *roll*) acquis par la centrale inertielle *InertiaCube3* au temps t_{IC3} à une image de la caméra thermique *Gobi 640 CL* acquise au temps t_{Gobi}, nous recherchons les données dont l'horodatage t_{IC3} est le plus proche de l'horodatage t_{Gobi}. Par cette méthode, une erreur d'horodatage ε est commise.

$$\varepsilon(l) = \min_{k}\{t_{Gobi}(l) - t_{IC3}(k)\} \tag{4.1}$$

L'horodatage de la $l^{ème}$ image de la caméra est noté $t_{Gobi}(l)$. L'ensemble de donnée (*yaw*, *pitch*, *roll*) correspond au temps $t_{IC3}(k)$ qui minimise l'équation (4.1) est associé à la $l^{ème}$ image aquise au temps $t_{Gobi}(l)$.

Nous avons vérifié notre système de synchronisation grâce au montage décrit sur la Figure 100. La caméra thermique filme la graduation d'une platine de rotation *Newport*. Les graduations sont visibles car elles possèdent une meilleure émissivité que le reste de la platine qui est un bon réflecteur.

Le carton permet de surélever la centrale inertielle par rapport aux matériaux de la platine susceptible de perturber les senseurs magnétiques de la centrale inertielle.

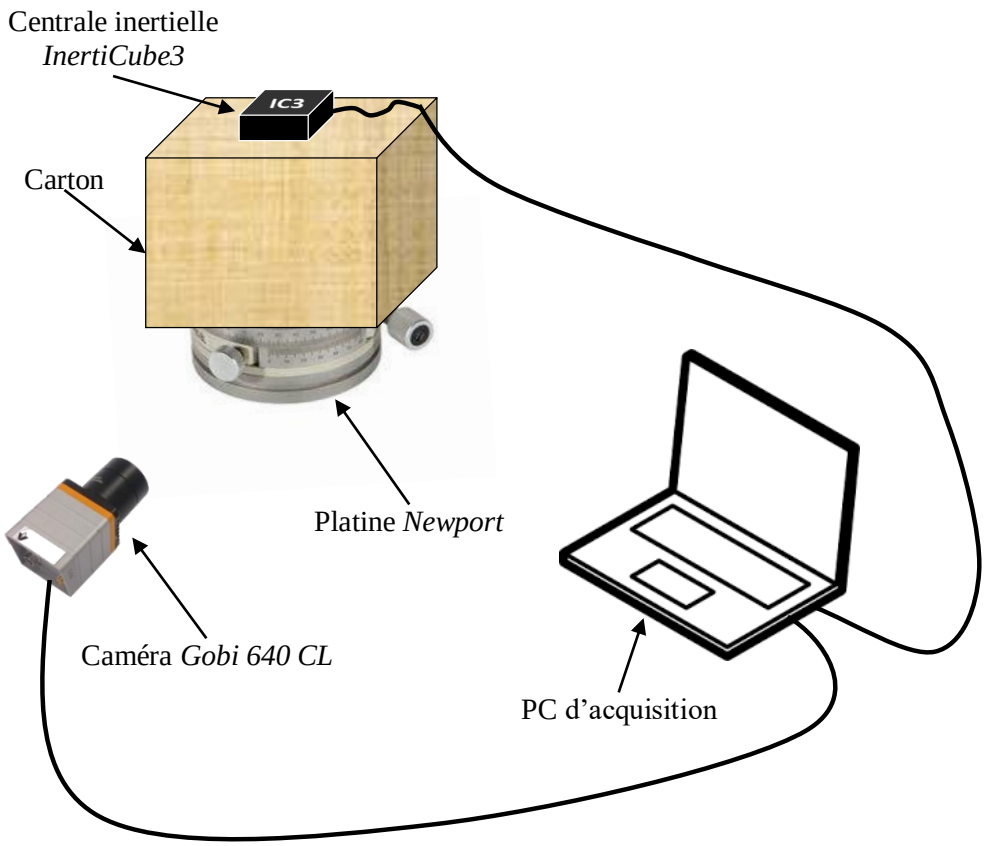


Figure 100. Procédure de vérification de la synchronisation entre la caméra *Gobi 640 CL* et la centrale inertielle *InertiCube3*.

La Figure 101 montre certaine images d'une séquence dans laquelle nous faisons pivoter la centrale inertielle solidaire du carton et de la platine de rotation. La graduation de la platine indique 240° (cela est tout à fait arbitraire) lors de de la première acquisition la caméra. L'angle *yaw* mesuré par la centrale inertielle est indiqué dans la zone de texte claire en haut de chaque image. Cet angle est à comparer avec la graduation de la platine de rotation. Dans la zone de texte, nous avons également indiqué l'erreur de temps ε. Nous initialisons manuellement l'angle *yaw* fournie par la centrale inertielle à 240° également à la première image acquise par la caméra (c'est pour cela que sur la première image de la Figure 101 l'angle *yaw* vaut exactement 240°).

Le mouvement appliqué est ample : de -40° à +65° autour de la valeur initiale de 240° pour simuler le type mouvement de la tête que nous souhaitons mesurer avec la centrale inertielle par la suite.

En considérant que la platine de rotation est l'étalon de cette expérience, nous constatons que l'erreur sur l'estimation de l'angle *yaw* par la centrale inertielle *InertiCube3* varie entre +/-1° à la vue des images de la Figure 101. Cet ordre de grandeur est compatible avec les données du fournisseur de la centrale inertielle. Cela montre également que notre méthode d'horodatage est suffisamment efficace.

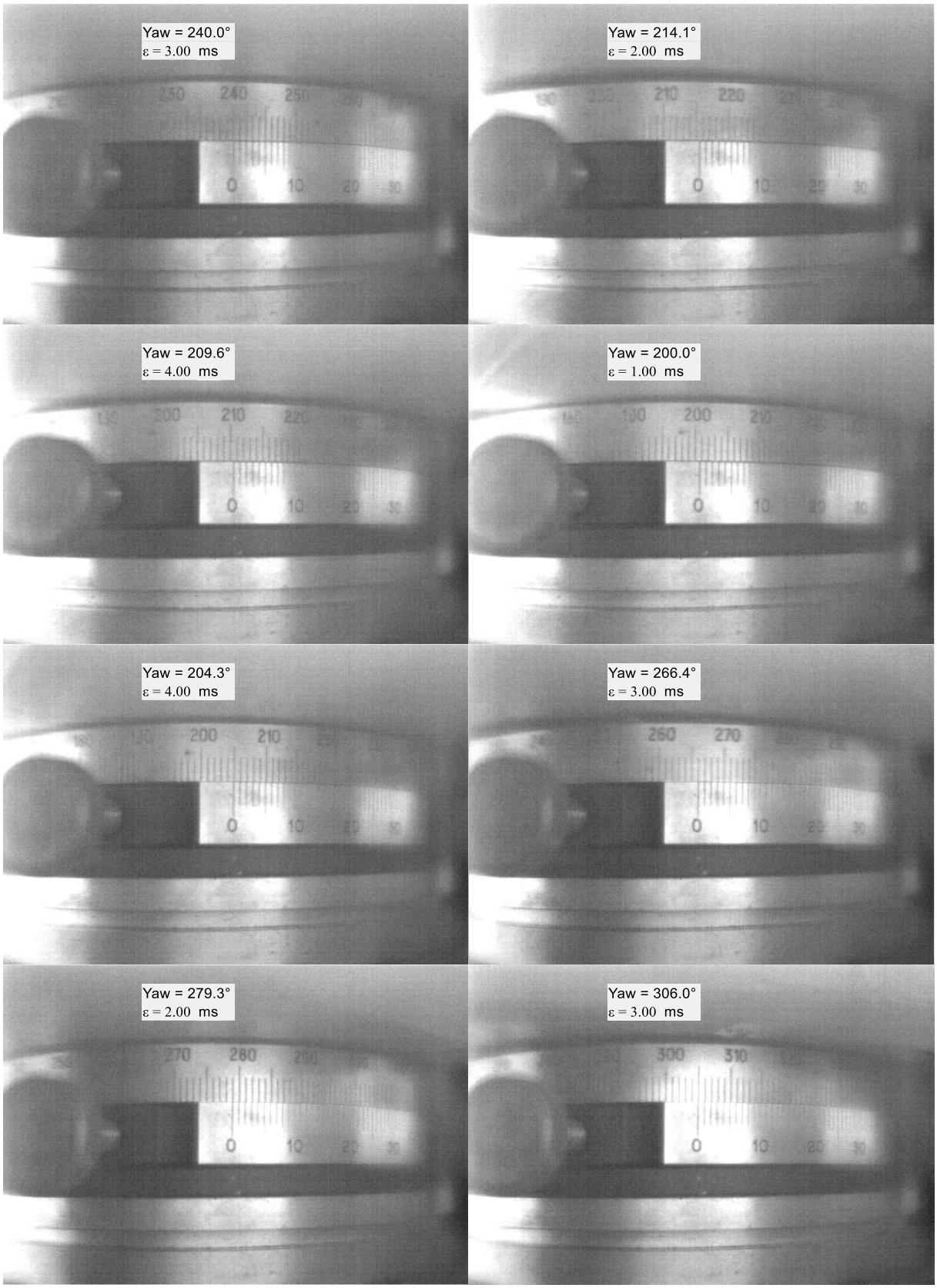


Figure 101. Images de contrôle de la synchronisation entre la centrale inertielle *InerticaCube3* et la caméra *Gobi 640 CL*. L'angle *yaw* indiqué dans la zone de texte correspond à la mesure de la centrale.

4.2.2. Un modèle simplifié d'un habitacle de véhicule

Nous souhaitons réaliser des séquences vidéos représentatives d'une situation réelle de conduite dans un véhicule. Nous aurions pu réaliser ces tests dans un habitacle réel mais nous avons choisis de créer un modèle simplifié d'un véhicule en laboratoire. La première raison qui justifie la création d'un modèle simplifié est la perturbation magnétique d'un habitacle réel sur les mesures de la centrale inertielle. La seconde raison est la facilité d'installation du matériel de mesure dans un espace moins contraint.

Un modèle simplifié d'un habitacle de Renault *Scenic* a été reproduit (cf. Figure 103). Neuf cibles ont été disposées à des endroits stratégiques :

- cible n°1 : milieu fenêtre latérale gauche,
- cible n°2 : rétroviseur latérale gauche,
- cible n°3 : milieu gauche du pare-brise (droit devant le conducteur),
- cible n°4 : compteur de vitesse,
- cible n°5 : rétroviseur central,
- cible n°6 : ordinateur de bord (radio),
- cible n°7 : milieu droit du pare-brise,
- cible n°8 : rétroviseur latéral droit,
- cible n°9 : fenêtre latérale droite.

La centrale inertielle est fixée sur un bandeau élastique porté par le participant. Lors d'un scénario typique, on demande au participant d'effectuer les mouvements suivants :

- « Regardez la cible n°3. » *on démarre le logiciel d'acquisition des données*.
- « Regardez les cibles numérotées de 1 à 9 dans l'ordre tout en restant fixé approximativement 3 s sur chaque cible. »
- « Après la cible 9, regardez à nouveau la cible n°3. » *on arrête l'acquisition des données*.

Nous laissons le participant libre de regarder les cibles comme il le souhaite. Il lui est donc possible d'orienter ses yeux et ses épaules.

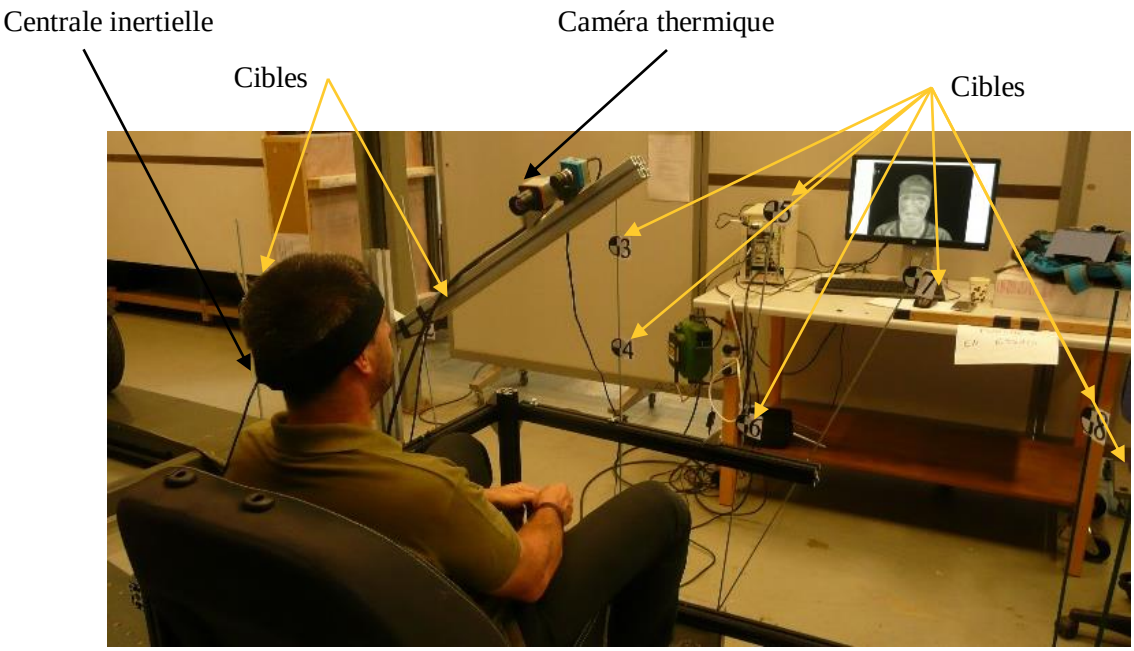


Figure 103. Un collaborateur de l'entreprise sur notre modèle simplifié d'habitacle du véhicule.

La fréquence d'acquisition est approximativement de 15 Hz. Un scénario typique dure environ 30 secondes. On enregistre approximativement 450 images. Les données typiquement relevées par la centrale inertielle lors d'un scénario sont représentés sur la Figure 102.

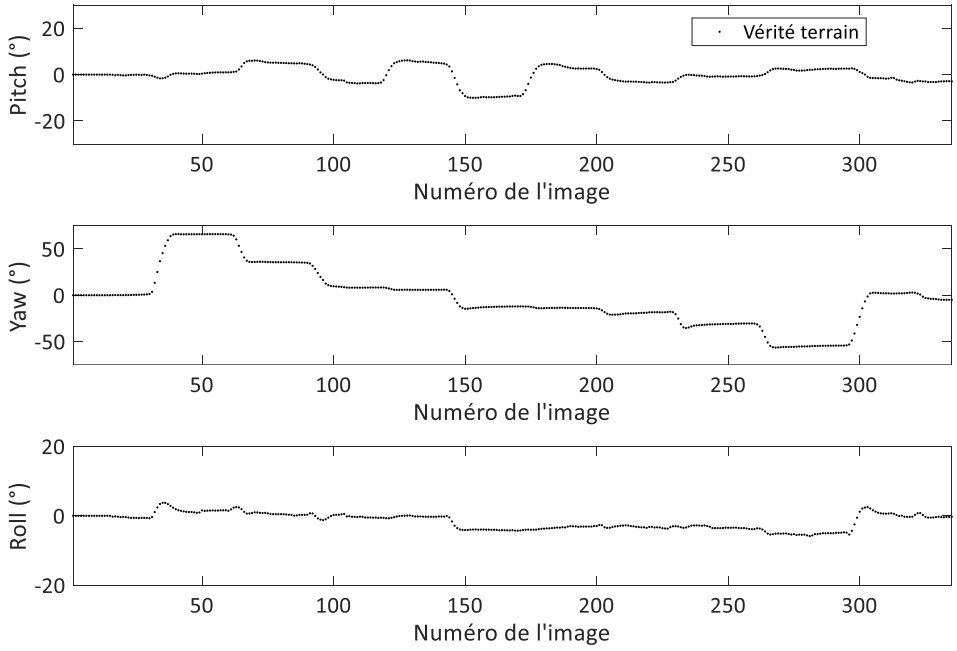


Figure 102. Données acquises par la centrale inertielle lors d'un scénario typique où le participant regarde successivement les neuf cibles du modèle simplifié du véhicule.

Les images brutes, c'est-à-dire sans aucune correction *NUC* sont enregistrées. Ainsi, nous nous laissons la possibilité de tester plusieurs méthodes de correction du bruit spatial fixe résiduel *BSFR*. La température de la caméra T_C est stable durant toute la durée de la vidéo et nous enregistrons environ 50 images de l'obturateur interne (*shutter*) en début de séquence.

Un panel de 20 participants a été recruté et nous a permis de créer une base de vidéos en imagerie thermique non-refroidie, avec la vérité terrain provenant de la centrale inertielle.

Remarque : avant que chaque participant ne s'installe dans le modèle simplifié du véhicule, nous enregistrons plusieurs images du fond (c'est-à-dire des images du modèle simplifié du véhicule sans le participant). Ces images peuvent être utilisée pour détecter (segmenter) le visage. Nous ne les utilisons pas dans notre algorithme d'estimation de la pose. En effet, dans le Chapitre 3 le plaquage de texture sur le maillage 3D est effectué manuellement. Nous traitons ainsi un problème de tracking et non de détection. Dans des perspectives d'automatisation de l'étape du plaquage de texture, la détection du visage sera indispensable. Ces images du fond pourraient être utilisées pour une segmentation robuste et rapide.

Il est à noter qu'il existe des méthodes de détection du visage en imagerie thermique qui ne sont pas basées sur la soustraction du fond. Par exemple, dans la référence [127], les auteurs proposent de détecter le visage dans une image thermique grâce à un seuillage automatique dont le seuil est déterminé par la méthode de maximisation de la variance interclasse (détaillé dans la référence [128]). Dans la référence [129] les auteurs précisent que d'après leurs expériences, lorsque la température ambiante est stable, la température de la peau varie dans un intervalle de température restreint, ce qui laisse entrevoir des possibilités de segmentation grâce au simple niveau des pixels. Dans la référence [130], les auteurs utilisent une caméra ayant subi un étalonnage radiométrique. Ainsi le niveau d'un pixel du visage doit être compris dans un intervalle indépendant de l'utilisateur et des conditions expérimentales (comme la température de la caméra). Ils attribuent alors une probabilité aux pixels de l'image d'appartenir à la peau du conducteur. Il est difficile de garantir que la détection du visage puisse fonctionner dans tous les cas d'usage grâce à ces méthodes basées sur une estimation de la température de la peau. Par exemple, si un véhicule est garé au soleil et qu'un élément de l'habitacle rayonne, il pourrait être confondu avec la classe normalement attribuée à la peau du conducteur. C'est pourquoi nous n'excluons pas d'utiliser les images du fond pour détecter automatiquement le visage pour automatiser le plaquage de texture.

4.2.3. Les figures de mérites pour évaluer l'estimation de la pose

Afin d'évaluer les performances des estimations des orientations, nous utiliserons l'erreur moyenne et l'écart type. Pour une image donnée, l'angle $\hat{\theta}$ est estimé par un algorithme de traitement d'images. L'angle θ est mesuré par la centrale inertielle. L'erreur d'estimation e commise pour un angle et pour une image donnée est :

$$e = \hat{\theta} - \theta \tag{4.2}$$

Si un algorithme de traitement d'images nous permet d'estimer chacun des angles *yaw*, *pitch* et *roll* alors on peut calculer l'erreur sur l'angle *yaw* notée e_{yaw}, l'erreur sur l'angle *pitch* notée e_{pitch} et l'erreur sur l'angle *roll* notée e_{roll}.

La moyenne de l'erreur m_e sur les N images d'une séquence vidéo est définie comme suit :

$$EM = \frac{1}{N}\sum_{i=1}^{N} \hat{\theta}_i - \theta_i \tag{4.3}$$

L'écart type de l'erreur σ_e commise sur les N images d'une séquence vidéo est défini comme suit :

$$S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i - m_e)^2} \tag{4.4}$$

L'erreur moyenne EM et l'écart type S peuvent donc être définis pour chacun des angles *yaw*, *pitch* et *roll*.

On définit une erreur maximale ξ comme étant l'erreur maximale parmi les angles estimés. Par exemple, pour la $i^{ème}$ image l'erreur maximale parmi les angles estimés est :

$$\xi(i) = \left|\max\left(e_{yaw}(i), e_{pitch}(i), e_{roll}(i)\right)\right| \tag{4.5}$$

Enfin, l'erreur maximale considérée seulement sur les angles *yaw* et *pitch*, est notée ξ'. Par exemple, pour la $i^{ème}$ image l'erreur maximale parmi les angles *yaw* et *pitch* est :

$$\xi'(i) = \left|\max\left(e_{yaw}(i), e_{pitch}(i)\right)\right| \tag{4.6}$$

4.3. Un algorithme d'estimation des angles *yaw* et *pitch* par une approche « problèmes inverses »

Nous disposons d'une base d'images de synthèse dont chacune des images est notée $\boldsymbol{g}_c$ avec $c\epsilon\{1, \ldots, 225\}$. A la $c^{ième}$ image de synthèse est associé un couple d'angle *(yaw,pitch)*. L'idée est de rechercher l'image de synthèse $\boldsymbol{g}_c$ qui est la plus proche de l'image réelle $\boldsymbol{Y}$ (c'est-à-dire l'image acquise par la caméra). On obtiendra ainsi une information sur le couple d'angle *(yaw,pitch)* de l'image réelle. Cette approche « problèmes inverses » nécessite de minimiser une fonction de coût notée $\mathcal{P}$ basée sur les moindres carrée pondérés entre l'image réelle et les images de synthèse.

Comme nous ne connaissons pas la position à priori du visage dans l'image réelle $\boldsymbol{Y}$, un vecteur décrivant le décalage entre la position du visage dans l'image réelle et la position du visage dans l'image de synthèse est à rechercher en même temps que les angles de la pose.

Le diagramme synoptique général de cet algorithme est récapitulé sur la Figure 104. Nous allons détailler dans cette section les hypothèses émises (section 4.3.1), la fonction de coût $\mathcal{P}$ utilisée (section 4.3.2), l'algorithme « global » (section 4.3.3) et les résultats obtenus (section 4.3.4).

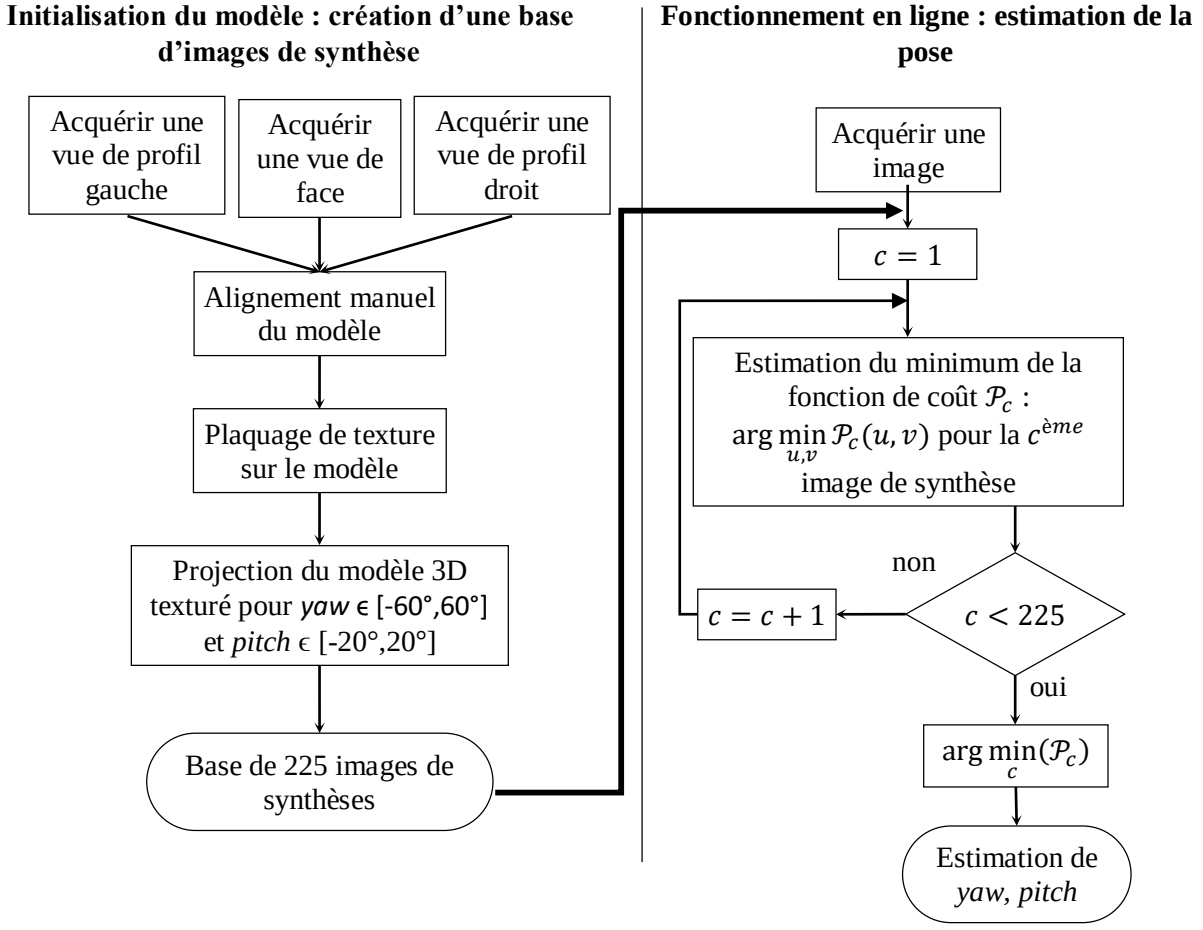


Figure 104. Diagramme synoptique de l'algorithme d'estimation des angles *yaw* et *pitch*.

4.3.1. Hypothèses simplificatrices

Nous faisons trois hypothèses qui réduisent le temps de calcul.

1. Nous considérerons qu'il n'y a pas d'offset entre l'image réelle $\boldsymbol{Y}$ et les images de synthèse $\boldsymbol{g}^c$. Cette hypothèse est réaliste si l'on considère que la température de caméra T_C varie peu entre l'instant de l'acquisition de la texture (utilisée pour créer les images de synthèses) et l'instant de l'acquisition de l'image réelle. Rappelons que pour que les réponses des pixels soient indépendantes de la température de la caméra T_C, soit nous acceptons de réaliser un étalonnage de longue durée, soit nous acceptons d'utiliser un élément de référence extérieur, comme un corps noir (cf. Chapitre 2, section 2.6).

2. Nous considérons qu'il n'y a pas de facteur d'échelle entre l'image réelle $\boldsymbol{Y}$ et les images de synthèse $\boldsymbol{g}^c$. Cette hypothèse signifie que l'on considère que la distance entre le conducteur et la caméra ne varie pas. En pratique cela n'est évidemment pas vérifié. La recherche d'un paramètre supplémentaire augmente

le temps de calcul. Pour contrer cela, l'accumulation temporelle des mesures combinée à l'hypothèse que le mouvement de la tête est continu et relativement lent est souvent utilisée. Ainsi, à l'aide de filtre de Kalman ou de filtre particulaire, il est possible de réduire le temps de calcul [123,131]. Notre objectif est simplement de valider la faisabilité d'un système d'estimation de la pose du visage du conducteur par imagerie thermique grâce à une approche « problèmes inverses ». Ainsi, seule la brique algorithmique de base, c'est-à-dire l'approche « problèmes inverses », sera testée.

3. Nous considérons que le conducteur ne réalise pas de rotation dans le plan. C'est-à-dire que nous considérons que l'angle *roll* est toujours nul. Si l'on se réfère à la Figure 102 et que l'on calcule la moyenne et l'écart type de l'angle *roll*, on obtient respectivement -1.9° et 1.67°. Dans la littérature scientifique, les algorithmes d'estimation de la pose ne sont pas aussi précis comme en témoigne la revue [58]. Cependant, bien que ce type de mouvement soit atypique, il devra être estimé correctement dans un produit final. Comme pour l'estimation du facteur d'échelle, l'estimation de l'angle *roll* augmente le temps de calcul. Les solutions pour contrer cela sont identiques à celle qui peuvent être mises en place pour estimer le facteur d'échelle, c'est-à-dire l'utilisation de l'accumulation temporelle des mesures et l'hypothèse d'un mouvement continu et relativement lent.

4.3.2. Approche « problèmes inverses » pour le recalage entre l'image de synthèse et l'image réelle

Nous testons exhaustivement chacune des 225 images de synthèse $\boldsymbol{g}$ pour identifier celle qui est la plus proche de l'image réelle $\boldsymbol{Y}$. Comme nous ne connaissons pas à priori la position du visage dans l'image

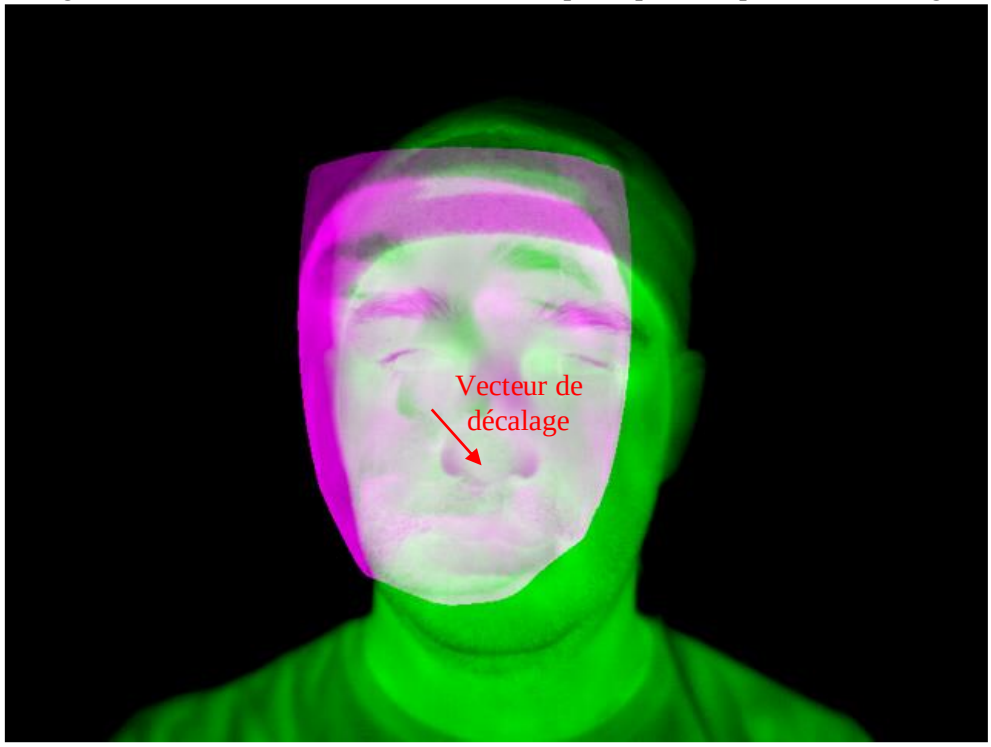


Figure 105. Illustration du décalage entre la position du visage dans l'image réelle et celle dans l'image de synthèse. L'image réelle est représentée en vert, l'image de synthèse est représentée en rose.

réelle, il est nécessaire de rechercher cette position à chaque image de synthèse testée. Le vecteur décalage décrit la différence de position du visage dans l'image réelle $\boldsymbol{Y}$ et dans l'image de synthèse $\boldsymbol{g}$ (cf. Figure 105).

Pour évaluer le décalage en position du visage nous cherchons le vecteur qui minimise l'écart quadratique pondéré entre l'image réelle et l'image de synthèse. Cette approche type « problèmes inverses » a déjà été utilisée dans de nombreux domaines. Par exemple, dans le cadre de l'holographie numérique, dans la référence [132], cette approche permet de localiser et d'estimer la taille de particules dont le modèle est connu. Dans notre cas, nous cherchons dans un premier temps seulement à localiser un modèle dans une image réelle. Nous considérons que les valeurs des pixels du visage de synthèse de coordonnées (i,j), notées g_{ij}, sont identique aux valeurs des pixels du visage réel de coordonnées (i,j) notées Y_{ij} à un bruit près ε_{ij} :

$$Y_{ij} = g_{ij} + \varepsilon_{ij} \; \forall i,j \tag{4.7}$$

Pour trouver la localisation $[u,v]$ du modèle, nous utilisons la fonction de coût $\mathcal{P}(u,v)$ suivante :

$$\mathcal{P}(u,v) = \sum_{ij} w_{uvij}\left(Y_{ij} - g_{uvij}\right)^2 \tag{4.8}$$

On note g_{ij} la valeur du pixel de la $i^{ème}$ ligne et de la $j^{ème}$ colonne de l'image de synthèse. On note g_{uvij} une image de synthèse décalée d'une valeur u selon l'axe vertical et décalée d'une valeur v selon l'axe horizontal. Comme hors du visage les pixels de l'image de synthèse valent arbitrairement 0, nous ne souhaitons pas que la fonction de coût prennent en compte ces pixels. Ainsi, nous multiplions l'écart quadratique par une matrice de poids $\boldsymbol{w}$ dont les éléments w_{ij} sont définis comme suit :

- $w_{ij} = 1$ où le visage est défini
- $w_{ij} = 0$ ailleurs

La matrice $\boldsymbol{w}$ peut être vue comme un masque binaire (cf. Figure 106). Enfin, la notation w_{uvij} décrit un décalage de $[u,v]^T$ du masque.

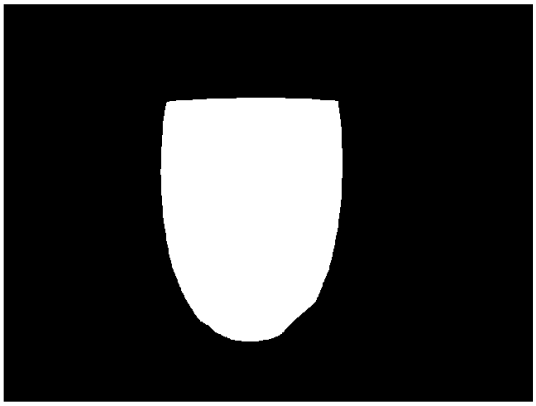


Figure 106. Masque $\boldsymbol{w}$ de pondération.

Si l'on développe l'équation (4.8) on obtient :

$$\mathcal{P}(u,v) = \sum_{ij} w_{uvij}.{Y_{ij}}^2 - 2\sum_{ij} w_{uvij}.Y_{ij}.g_{uvij} + \sum_{ij} w_{uvij}.\left(g_{uvij}\right)^2 \tag{4.9}$$

Le premier terme de l'équation (4.9) peut s'écrire sous la forme d'un produit de corrélation en fonction des lignes et des colonnes de l'image de la façon suivante :

$$\sum_{ij} w_{uvij}.{Y_{ij}}^2 = \sum\nolimits_i \sum\nolimits_j w(i-u, j-v).\left(Y(i,j)\right)^2 \tag{4.10}$$

Pour évaluer la fonction de coût, il est nécessaire de réaliser ce calcul pour l'ensemble des vecteurs$[u,v]^T$, c'est-à-dire pour l'ensemble des 480×640 pixels de l'image. Cette charge de calcul est conséquente. En s'inspirant de la référence [132] nous réduisons le temps de calcul effectuant le produit de corrélation dans l'espace de Fourier.

On peut également réécrire le second terme de l'équation (4.9) sous forme d'un produit de corrélation. En effet, le terme $w_{uvij}g_{uvij}$ est en fait une seule fonction réelle qui dépend de i et j et qui est décalée du même vecteur $[u,v]^T$. On la notera $wg(i-u, j-v)$. Ainsi,

$$\sum_{ij} Y_{ij} w_{uvij} g_{uvij} = \sum\nolimits_i \sum\nolimits_j Y(i,j).wg(i-u, j-v) \tag{4.11}$$

Enfin le troisième terme de l'équation (4.9) est constant et se calcul grâce à une multiplication de matrice et une somme.

Remarque : l'utilisation d'un masque est contraignante car elle nécessite le calcul de deux produits de corrélation au lieu d'un seul. Il est cependant indispensable car les images thermiques sont peu texturées et le produit de corrélation entre le visage de synthèse et une image réelle peut prendre des valeurs importantes dans le fond de l'image réelle.

Numériquement, un produit de corrélation peut être calculé à l'aide de trois transformés de Fourier (*FFT*). En effet, un produit de corrélation dans l'espace réel est une multiplication dans l'espace de Fourier. Deux FFTs sont donc utilisées pour cela. Une troisième permet d'exprimer le résultat dans l'espace réel. Pour l'estimation des coordonnées (u_c, v_c) pour une image de la base de synthèse, le calcul de la fonction de coût est effectué à partir de six *FFTs*.

On note $[u_0, v_0]^T$ le centre de l'image de synthèse. Etant donné que nous utilisons une caméra *VGA* (c'est-à-dire au format 480×640 pixels), $u_0 = 241$ et $v_0 = 321$. Le décalage entre la position du visage dans l'image réelle et la position du visage dans la $c^{ème}$ image de synthèse est $[u_c - u_0, v_c - v_0]^T$ avec les coordonnées (u_c, v_c) qui minimisent $\mathcal{P}(u,v)$:

$$\mathcal{P}_c(u_c, v_c) = \arg\min_{u,v} \mathcal{P}_C(u,v) \tag{4.12}$$

La Figure 107 illustre la carte de l'écart quadratique pondéré pour une image réelle de face et l'image de synthèse crée à partir du couple ($yaw = 0, pitch = 0$), c'est-à-dire les deux images représentées en superposition sur la Figure 105. Le minimum est situé aux coordonnées (u_c, v_c).

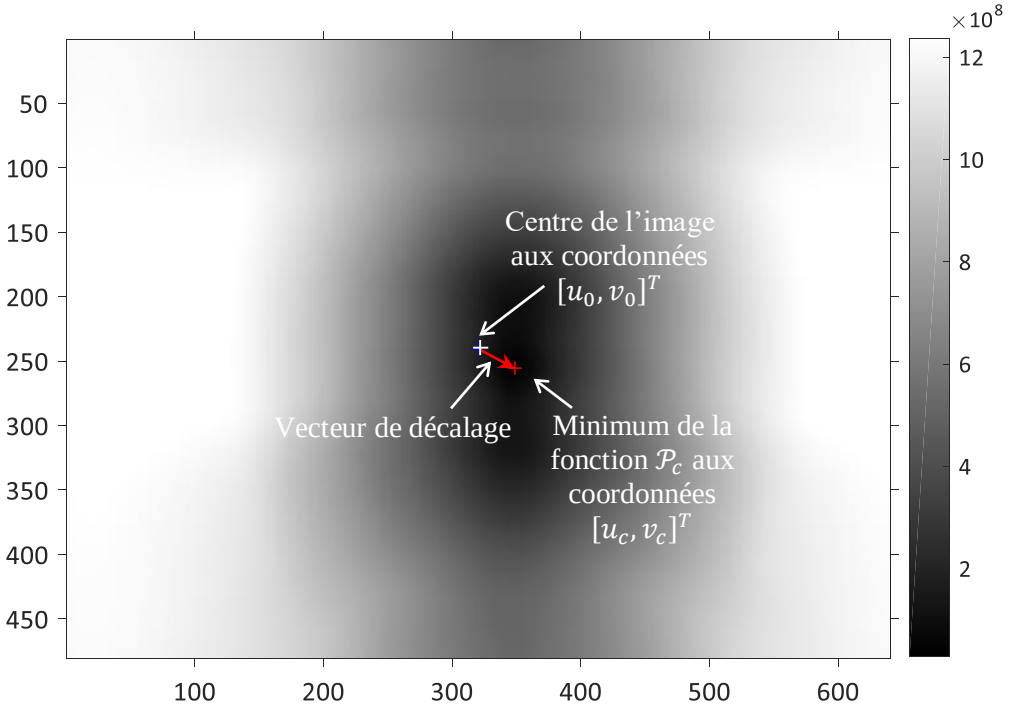


Figure 107. Carte des valeurs de l'écart quadratique pondéré : $\mathcal{P}_C(u, v)$.

Nous translatons l'image de synthèse grâce au vecteur de décalage $[u_c - u_0, v_c - v_0]^T$. On obtient la Figure 108.

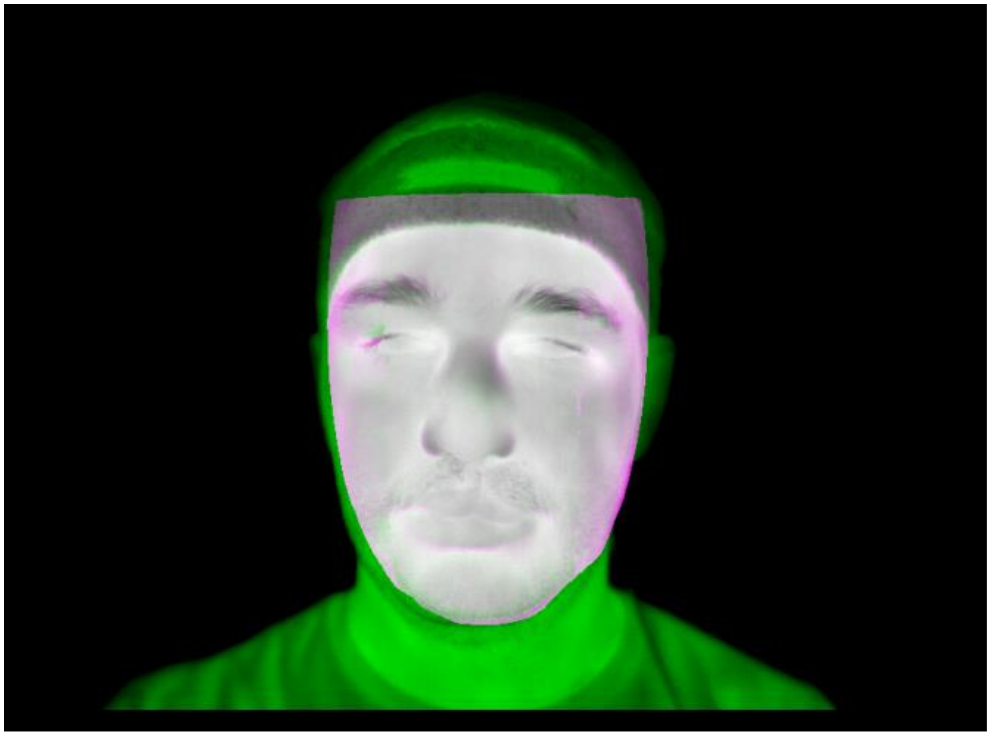


Figure 108. Illustration de l'alignement de l'image de synthèse avec l'image réelle. L'image réelle est représentée en vert, l'image de synthèse est représentée en rose.

4.3.3. Algorithme « global »

Comme cela est expliqué à le section précédente, pour chacune des 225 images de synthèse, nous recherchons la position (u_c, v_c) qui minimise l'écart quadratique pondéré $\mathcal{P}(u, v)$. Ensuite, nous sauvegardons les valeurs $\mathcal{P}_c(u_c, v_c)$, u_c et v_c pour $c \in \{1, \dots, 225\}$, c'est-à-dire pour toutes les images de synthèse de la base. Puis nous recherchons l'image de synthèse qui minimise $\mathcal{P}_c(u_c, v_c)$ parmi les 225.

On cherche la $c^{ème}$ $i$mage de synthèse telle que :

$$\mathcal{P}_c(u_c, v_c) = \arg\min_c \mathcal{P}_c(u_c, v_c) \text{ avec } c \in \{1, \dots, 225\} \tag{4.13}$$

La Figure 109 est un diagramme synoptique qui illustre (i) la recherche de la position qui minimise l'écart quadratique pondéré entre une image réelle et une image de synthèse (étape qui est réalisée pour chacune des 225 images de synthèse) et (ii) la recherche de l'image de synthèse qui minimise l'écart quadratique pondéré.

Remarque : il est indispensable de veiller à ce que la dynamique des images de synthèses et des images réelles ne soit pas modifiée par un ajustement automatique, ou une normalisation.

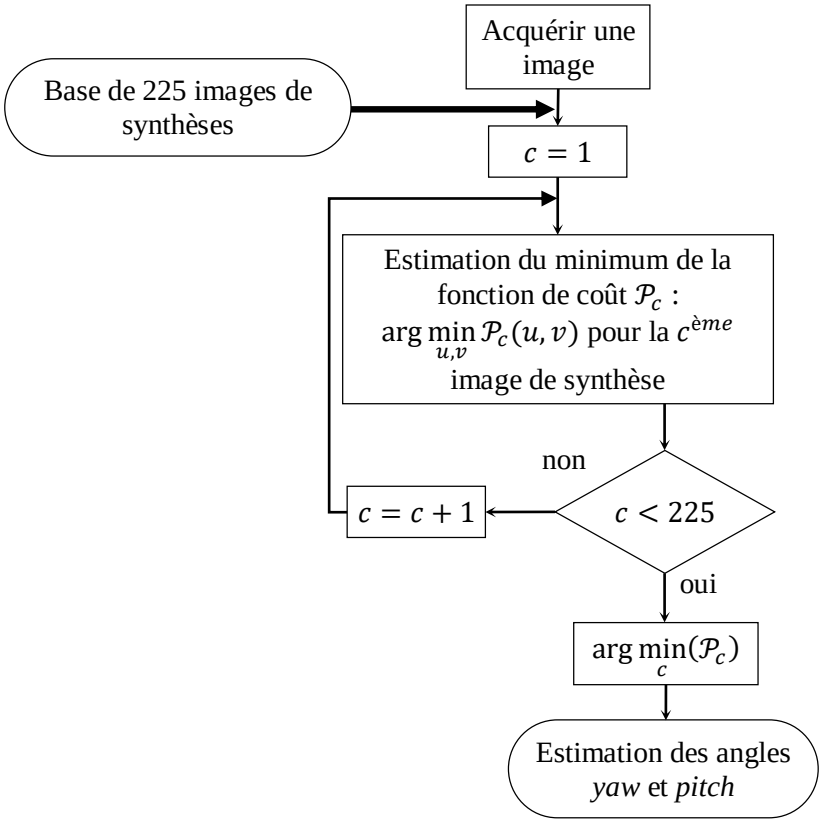


Figure 109. Diagramme synoptique de la recherche de l'image de synthèse qui minimise l'écart quadratique pondéré avec l'image réelle. Pour chacune des 225 images de synthèse, on recherche une première fois la position qui minimise l'écart quadratique pondéré avec l'image réelle. Puis, l'image de synthèse qui possède le plus petit écart quadratique pondéré avec l'image réelle est déclarée comme étant celle qui décrit le mieux l'image réelle. Le couple d'angle (*yaw,pitch*) sont les paramètres recherchés.

La Figure 110 illustre plus particulièrement le choix de l'image de synthèse après de l'optimisation de sa position.

La Figure 111 représente l'image réelle (en haut à gauche) et l'image de synthèse (en haut à droite) qui minimise la fonction de coût $\mathcal{P}_c(u_c, v_c)$ exprimée par l'équation (4.13). Les valeurs de $\mathcal{P}_c(u_c, v_c)$ pour chacune des images de synthèse sont représentées sur le graphique en bas de la Figure 111. La Figure 112 représente exactement les mêmes éléments mis à part que le maillage 3D utilisé est l'ellipsoïde.

Finalement, puisque nous testons exhaustivement les 225 images de synthèse de la base, pour une image réelle nous calculons 225×6=1350 *FFTs*.

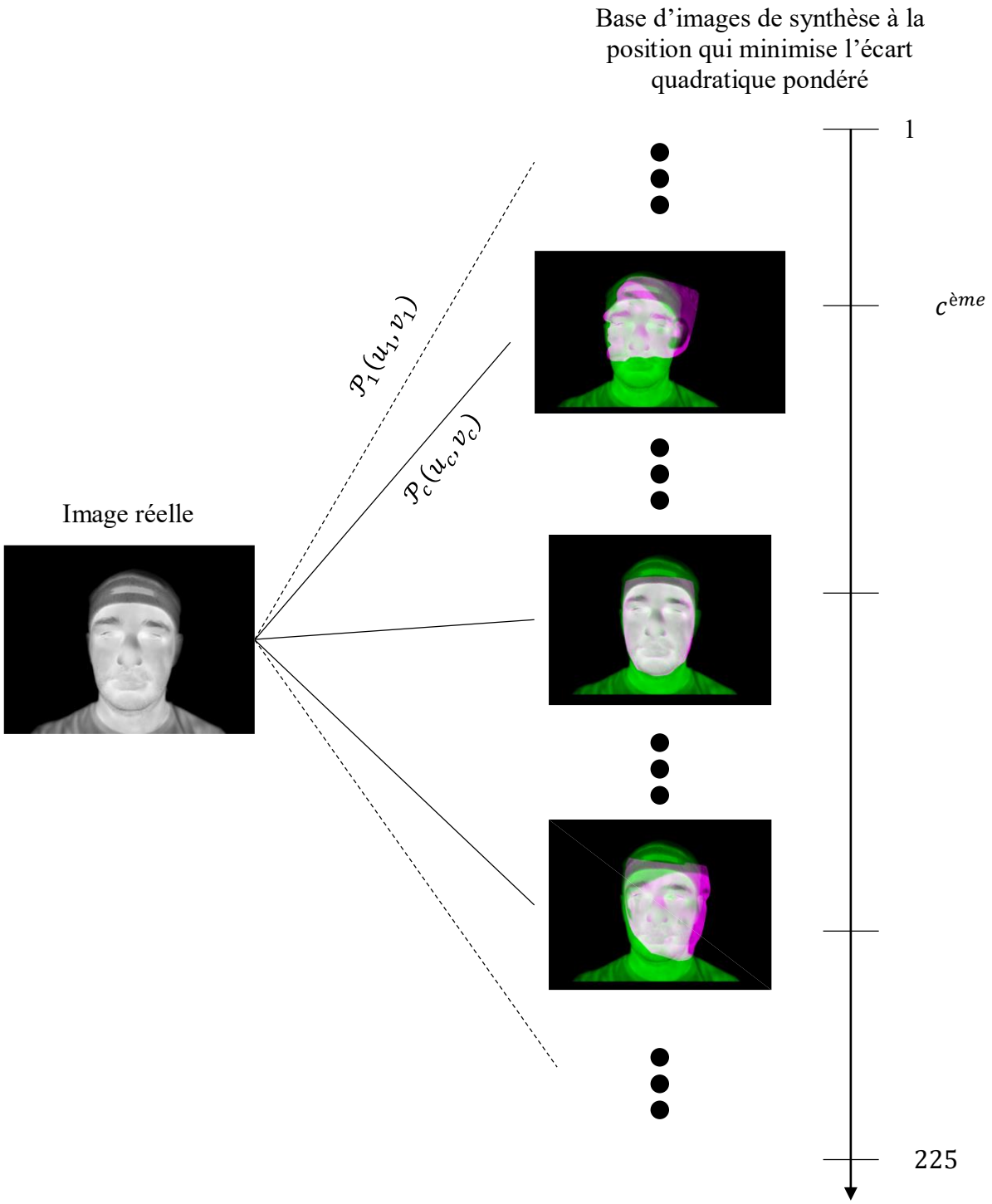


Figure 110. Illustration du choix de l'image de synthèse.

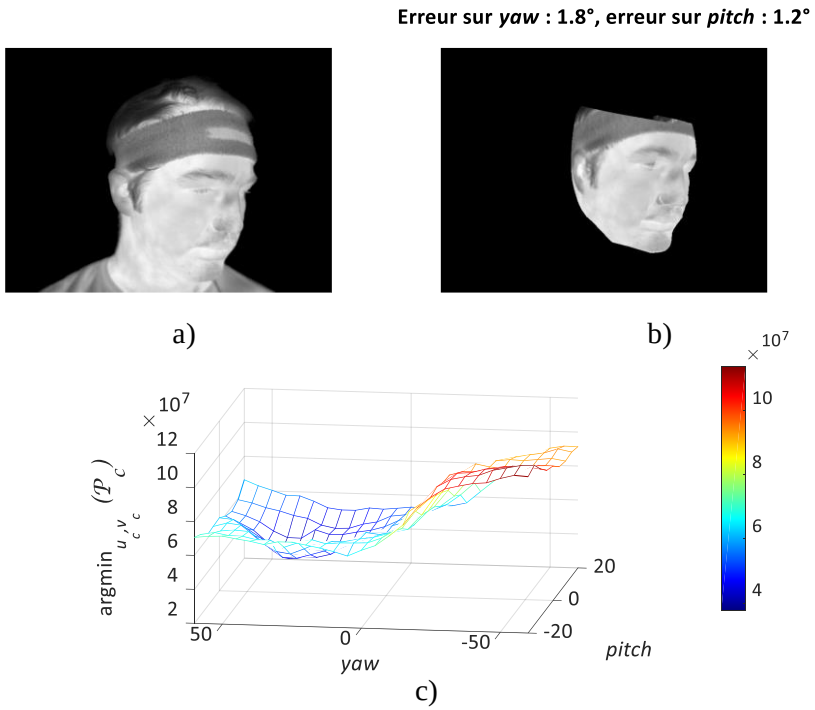


Figure 111. Illustration du fonctionnement de l'algorithme avec le maillage 3D précis. a) Image réelle, b) image de synthèse obtenue avec le maillage précis qui minimise la fonction de coût $\mathcal{P}_c(u_c, v_c)$ exprimée par l'équation (4.13), c) les valeurs de $\mathcal{P}_c(u_c, v_c)$ pour chacune des images de synthèse.

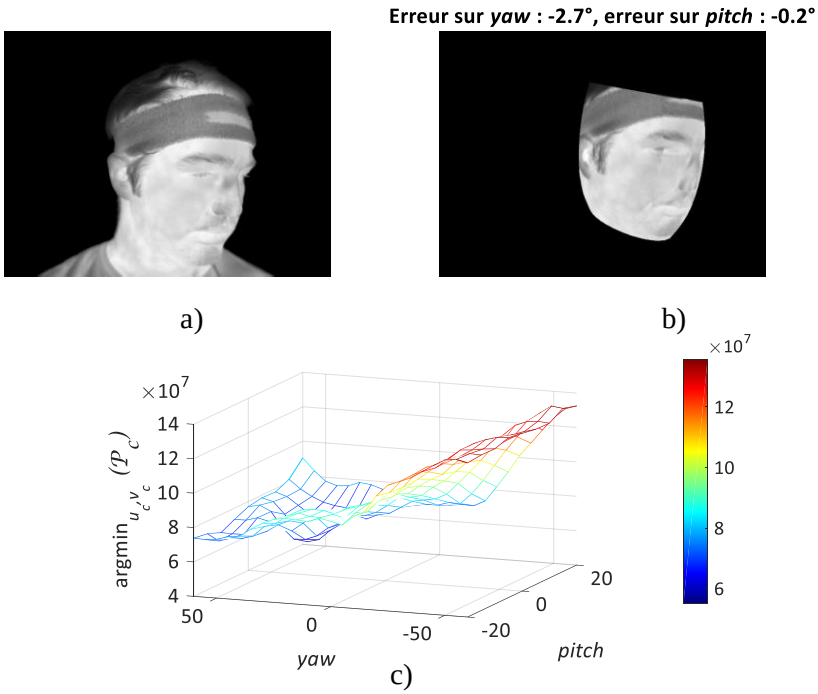


Figure 112. Illustration du fonctionnement de l'algorithme avec le maillage 3D ellipsoïdal. a) Image réelle, b) image de synthèse obtenue avec le maillage précis qui minimise la fonction de coût $\mathcal{P}_c(u_c, v_c)$ exprimée par l'équation (4.13), c) les valeurs de $\mathcal{P}_c(u_c, v_c)$ pour chacune des images de synthèse.

Estimation des angles yaw et pitch avec une meilleure précision que la pas d'échantillonnage :

La précision de l'estimation des angles *yaw* et *pitch* est limitée par le pas d'échantillonnage choisi par l'expérimentateur. Dans notre implémentation, nous rappelons que nous avons 225 images de synthèse d'un visage orienté selon un couple d'angle *(yaw,pitch)* spécifique avec *yaw* allant de -60° à +60° par pas de 5° et *pitch* allant de -20° à + 20° par pas de 5°.

La fonction $\mathcal{P}$ peut être vue comme une fonction qui dépend des deux variables *yaw* et *pitch*, c'est-à-dire la fonction $\mathcal{P}(yaw, pitch)$. Nous cherchons le minimum de cette fonction dans un espace échantillonné avec un pas de 5°. Afin d'améliorer l'estimation de $[u, v]$, nous effectuons une régression polynomiale autour de la valeur minimum. Nous simplifions cette régression 2D en deux régressions 1D (cf. annexe C). Nous ferons référence à cette technique classique grâce au terme *parabolic estimation* que l'on trouve dans la référence [133].

4.3.4. Limitation et solutions

Nous avons considéré jusqu'à présent que l'image de synthèse et l'image réelle étaient identiques à un bruit près (cf. équation (4.7)). Cette hypothèse est vérifiée si la température de la caméra T_C n'évolue pas. Dans le cas contraire un offset global et un gain global modifient l'image.

Expérience portant sur l'influence d'un offset β :

Nous nous sommes attardés sur l'influence d'un offset global noté β. Nous avons souhaité évaluer quantitativement, sur un exemple, à partir de quelle valeur était-il possible de faire l'hypothèse qu'il n'y avait pas d'offset global β entre l'image réelle et les images de synthèse. Pour cela, nous avons répété l'estimation des angles *yaw* et *pitch* sur une unique image réelle pour un offset β qui varie entre 0 et 380 *Adu* par pas de 20 *Adu* et nous avons relevé l'erreur maximale ξ' (cf. équation (4.6)) sur les angles *yaw* et *pitch* (cf. Figure 113). Cette expérience montre qu'un offset supérieur à **320 *Adu*** (sur 16 bits) entre l'image réelle et les images de synthèse augmente drastiquement l'erreur ξ'.

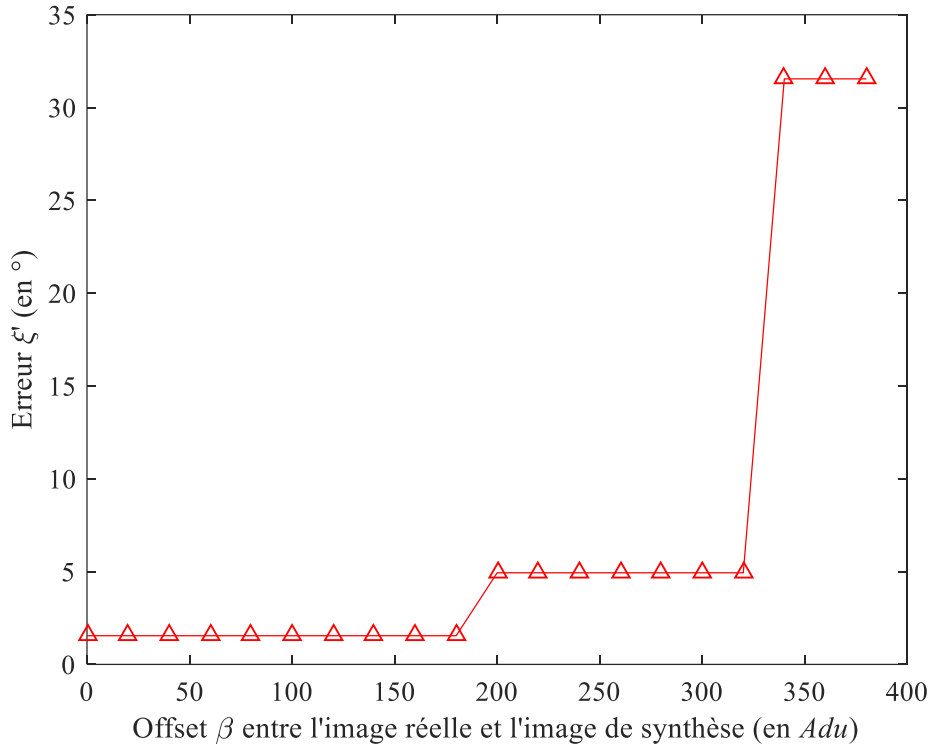


Figure 113. Erreur maximale sur les angles *yaw* et *pitch* en fonction d'un offset global β entre l'image réelle et les images de synthèse.

Un étalonnage radiométrique est possible comme nous en avons discuté au Chapitre 2 dans la section 2.6. Il permet de limiter les fluctuations des réponses des pixels. Plus précisément, dans un intervalle de température de la caméra de 25.2 à 54.9°C, 99.8% des pixels varient dans un intervalle inférieur à ±320 *Adu*.

Si un tel étalonnage n'est pas possible, il faut considérer un premier ordre un offset entre l'image de synthèse et l'image réelle et au deuxième ordre un offset et un gain. Regardons ce qu'il est nécessaire de mettre en œuvre pour considérer un offset β entre l'image de synthèse et l'image réelle :

$$Y_{ij} = g_{ij} + \varepsilon_{ij} + \beta \tag{4.14}$$

La fonction de coût à minimiser est la suivante :

$$\mathcal{P}(u,v) = \sum_{ij} w_{uvij}\left(Y_{ij} - g_{uvij} - \beta\right)^2 \tag{4.15}$$

Le développant de l'équation (4.15), effectué dans l'annexe C, fait apparaître trois produits de corrélations, ce qui en fait un de plus que dans la situation où l'on ne considère pas d'offset entre l'image de synthèse et les l'image réelle. Numériquement, ce calcul nécessite neuf *FFTs* (contre six dans la situation précédente). Pour illustrer l'apport de la prise en compte d'un offset β dans la fonction de coût représentée

par l'équation (4.15), reprenons « l'expérience portant sur l'influence d'un offset β » menée ci-dessus. Les réultats sont illustrés sur la Figure 114.

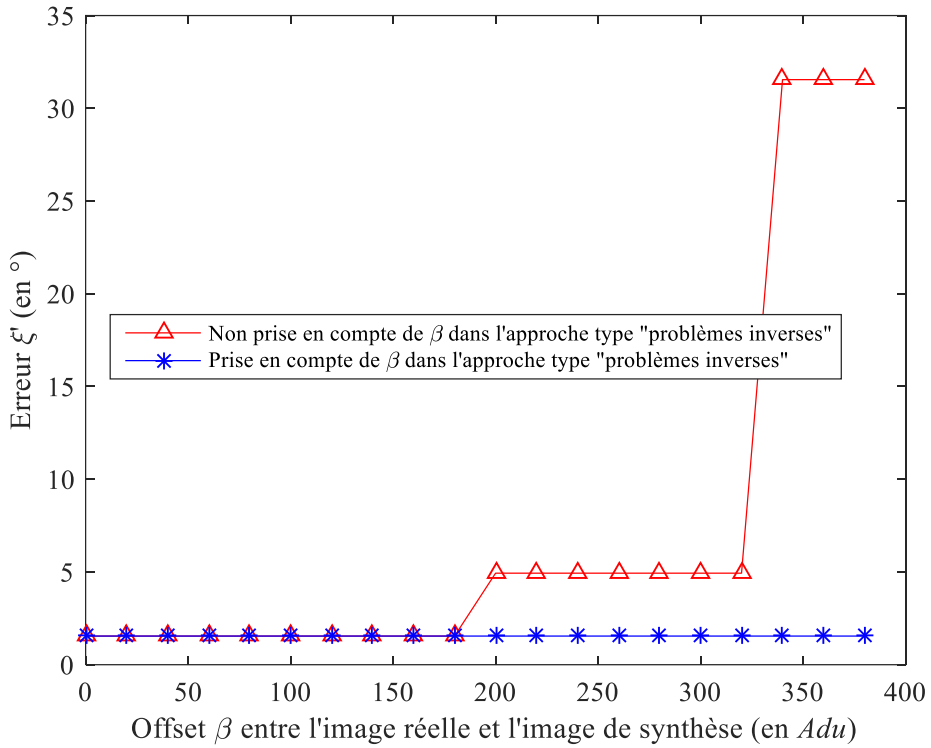


Figure 114. Erreur maximale sur les angles *yaw* et *pitch* en fonction d'un offset global β entre l'image réelle et les images de synthèse. La fonction de coût de l'approche type « problèmes inverses » ne prend pas en compte l'offset β (courbe rouge avec des triangles). Elle prend en compte β (courbe bleue avec des $*$).

Concluons donc sur le fait qu'une caméra bénéficiant d'un étalonnage radiométrique permet de réduire le temps de calcul dans le cas d'une approche « problèmes inverses » pour l'estimation de l'orientation d'un visage tel que cela a été implémenté dans cette section.

4.3.5. Résultats

Nous considérons un cas où la caméra est étalonnée radiométriquement. En pratique, pour faire cela simplement, nous utilisons des images du visages acquises à une température de la caméra T_{C1} pour créer l'avatar thermique et donc les images de synthèse. Ensuite, nous traitons des images réelles acquises à cette même température T_{C1}. Les résultats obtenus avec un maillage 3D précis (*Artec3D*) sont représentés sur la Figure 115. Ceux obtenus avec un maillage 3D ellipsoïdal sont reportés sur la Figure 116. Nous représentons sur ces deux figures la vérité terrain issue de la centrale inertielle en noir, les estimations avant le raffinement grâce à la méthode *parabolic estimation* en rouge et les estimations après le raffinement en bleu.

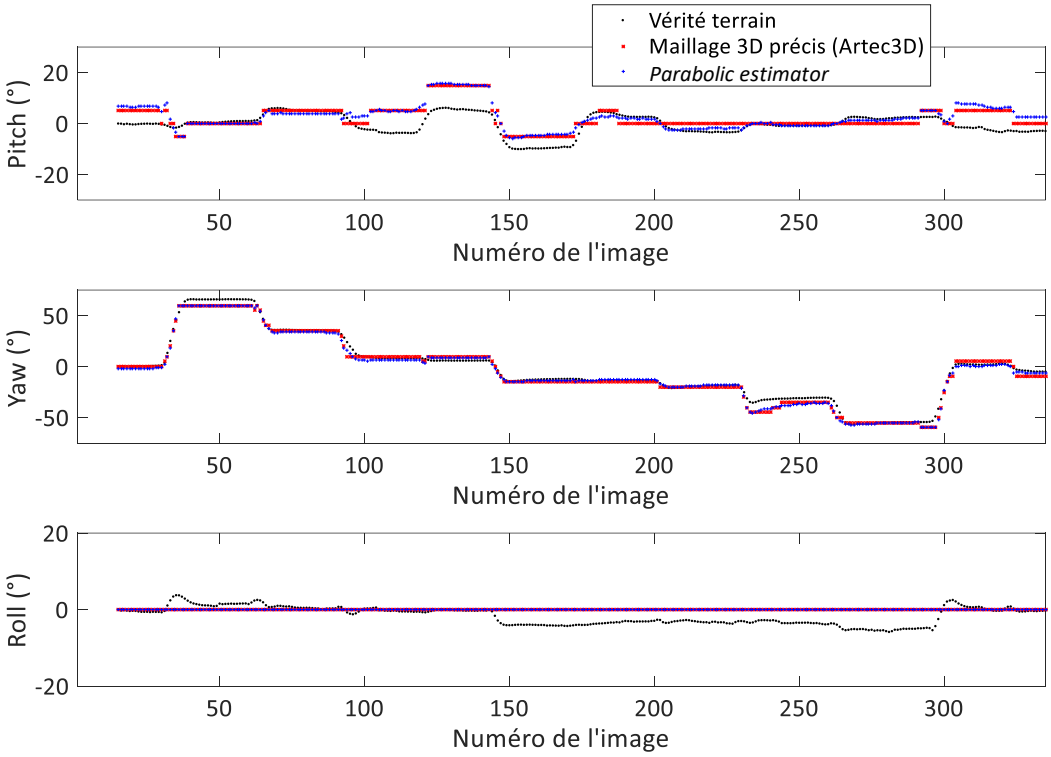


Figure 115. Résultats obtenus avec un maillage 3D précis (*Artec 3D*).

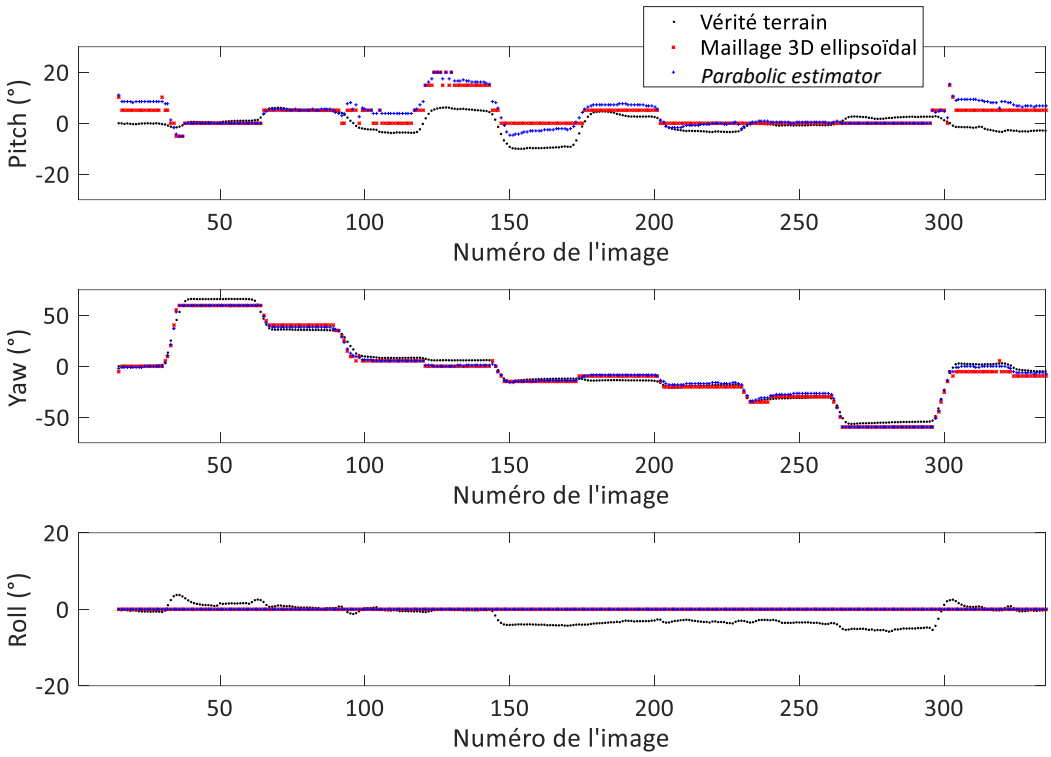


Figure 116. Résultats obtenus avec un maillage 3D ellipsoïdal.

L'angle *roll* n'étant pas estimé dans notre implémentation, il a été arbitrairement fixé à zéro dans la Figure 115 et dans la Figure 116. Une possibilité pour l'estimer consisterait à ajouter dans la base des images de synthèse avec un angle *roll* différent de 0 comme nous l'avons évoqué dans les hypothèses simplificatrices (section 4.3.1).

Les Tableau 8 et Tableau 9 répertorient l'erreur moyenne EM (sur les toutes images de la séquence vidéo à traiter) pour les deux angles estimés (*yaw* et *pitch*), ainsi que l'écart type de l'erreur S. Nous présentons les résultats sous la forme $EM \pm S$ comme cela est souvent le cas dans la littérature qui traite ce genre de problème. Les résultats sont présentés avec et sans la méthode *parabolic estimator*.

Tableau 8. Erreur moyenne et écart type de l'erreur sur l'estimation des angles *yaw* et *pitch*. Ces résultats ont été obtenus avec le maillage 3D précis (*Artec 3D*).

Méthode *parabolic estimator*	**Résultats des estimations sur les rotations** $(EM \pm S)$		
	Erreur sur *pitch*	**Erreur sur *yaw***	**Erreur sur *roll***
non	-2.3° ± 3,8°	1.6° ± 3.7°	/
oui	-2.7° ± 4,0°	2.1° ± 3.1°	/

Tableau 9. Erreur moyenne et écart type de l'erreur sur l'estimation des angles *yaw* et *pitch*. Ces résultats ont été obtenus avec le maillage 3D ellipsoïdal.

Méthode *parabolic estimator*	**Résultats des estimations sur les rotations** $(EM \pm S)$		
	Erreur sur *pitch*	**Erreur sur *yaw***	**Erreur sur *roll***
non	-3.3° ± 4,3°	2.0° ± 3.8°	/
oui	-4.1° ± 4,7°	0.9° ± 3.7°	/

On peut noter que l'estimation de l'angle *pitch* semble moins précise que l'estimation de l'angle *yaw*. Si l'on s'attarde sur les mauvaises estimations de l'angle *pitch* sur les Figure 115 et Figure 116, et que l'on regarde l'angle *yaw* correspondant à la même image réelle, on s'aperçoit que les erreurs sur *pitch* surviennent majoritairement lorsque l'angle *yaw* est proche de zéro. Nous proposons d'expliquer cela par le fait que lors du plaquage de texture de la partie de face du visage, l'orientation *pitch* donnée au maillage 3D est entachée d'une erreur. Celle-ci se répercute logiquement sur les estimations de l'angle *pitch* lorsque l'angle *yaw* est proche de zéro.

On s'aperçoit que l'utilisation d'un maillage 3D ellipsoïdal réduit la précision des estimations des angles *pitch* et *yaw*. Cependant, cette détérioration reste faible, elle est inférieure à 1° en moyenne. Du point de vu de la robustesse de l'algorithme aux variations inter-individus des formes 3D des visages, cette constatation peut être interprétée comme une bonne nouvelle. Elle signifie qu'un maillage 3D générique permet d'obtenir des résultats raisonnables, et donc, qu'un maillage 3D précis personnalisé à l'utilisateur, ne sera pas forcément nécessaire.

Enfin, concernant la méthode d'estimation *parabolic estimator*, on s'aperçoit, aux vus des deux tableaux, et des figures qu'elle n'améliore pas toujours la situation. La raison de cela est que le modèle que nous utilisons n'est pas aussi réaliste que nous le souhaitons (notamment à cause d'un mauvais alignement du maillage 3D avec l'image lors de la phase de plaquage de texture). De plus, cette méthode est adaptée à la détection de pics symétriques, ce qui n'est pas nécessairement le cas.

La Figure 117 et la Figure 118 représentent des images réelles de la séquence auxquelles nous avons superposé les images de synthèse choisies, aux positions qui minimisent l'écart quadratique pondéré.

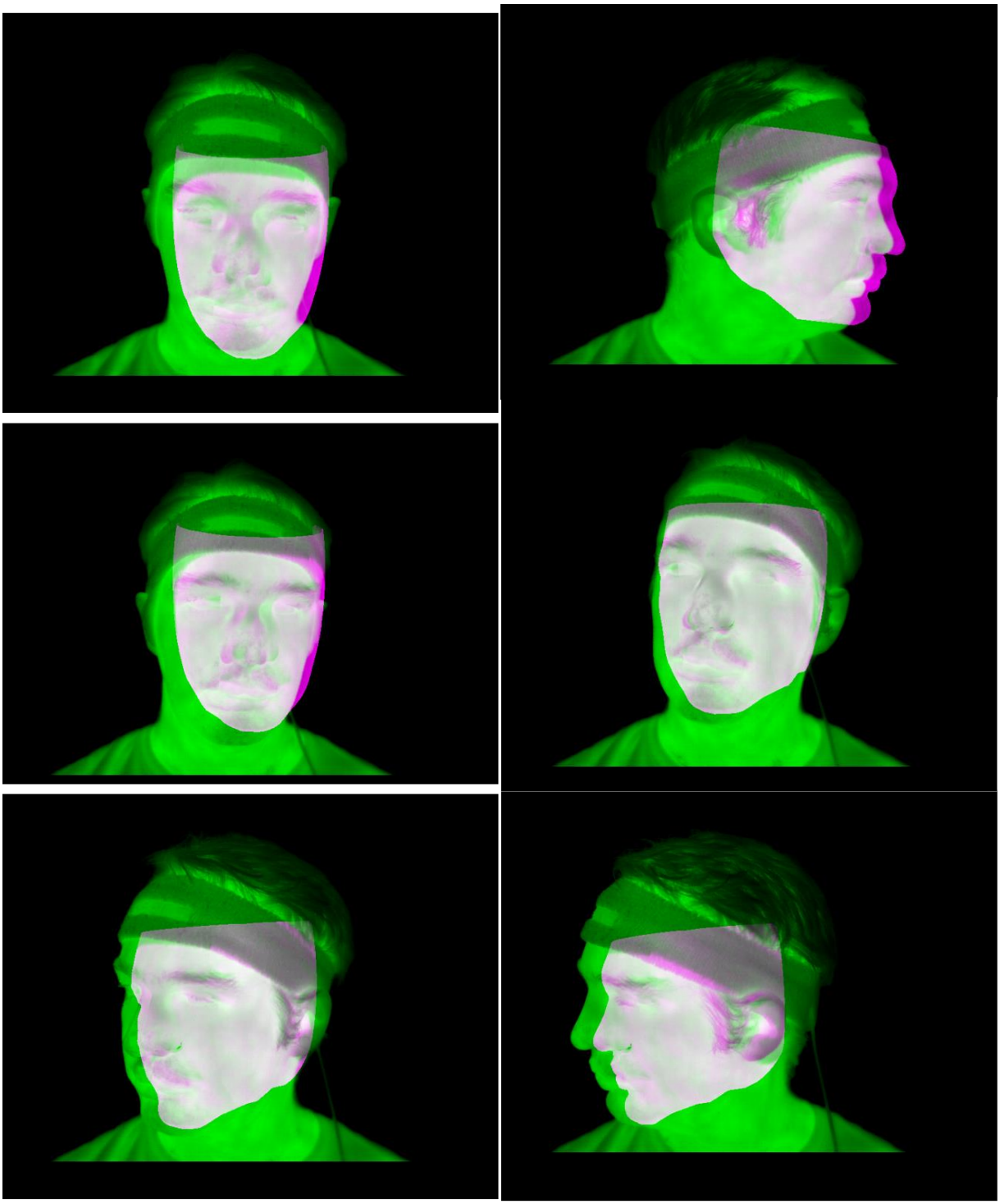


Figure 117. Illustration des résultats obtenus avec le maillage 3D précis. L'image réelle (en vert) et l'image de synthèse (en rose) sont superposées.

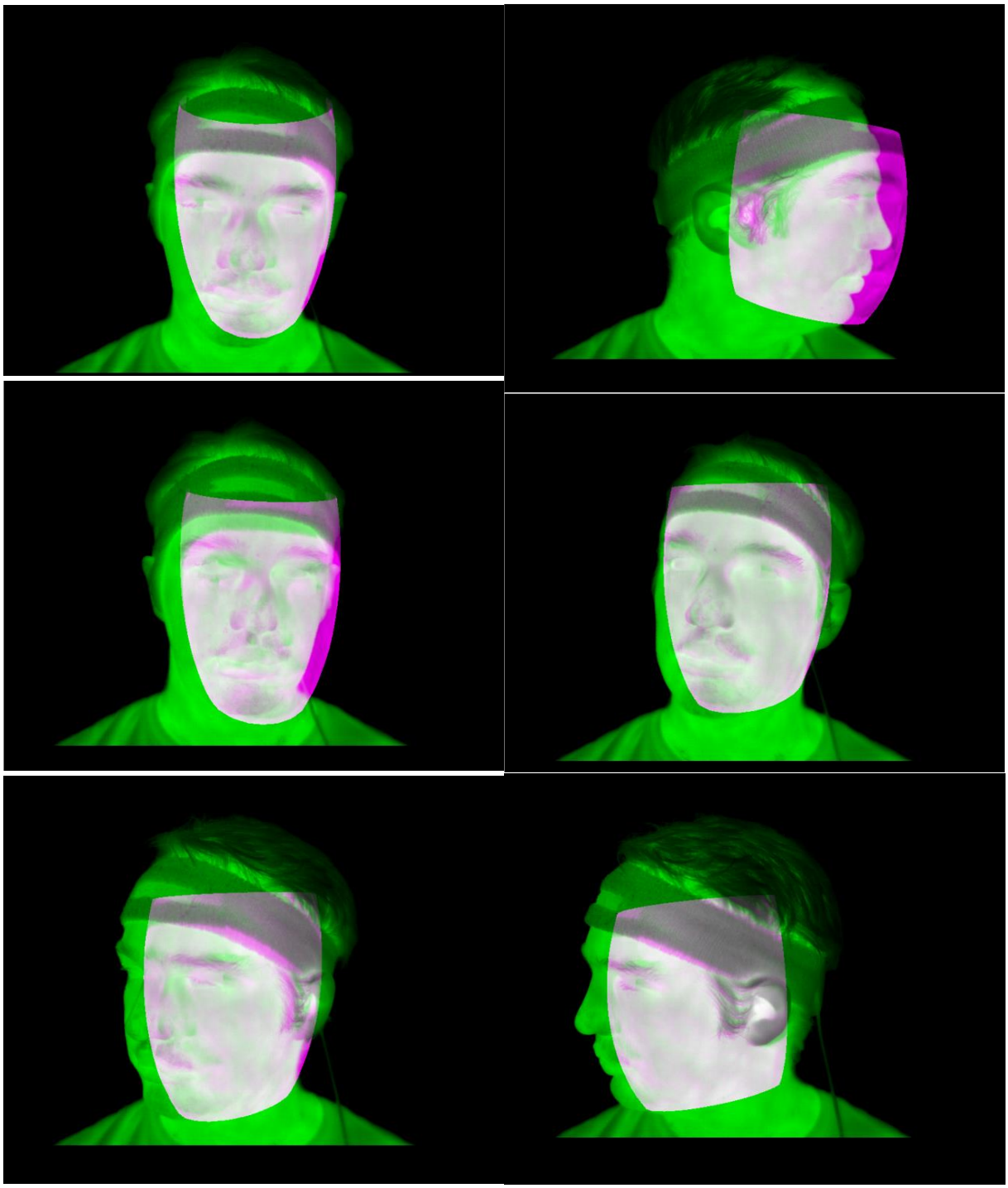


Figure 118. Illustration des résultats obtenus avec le maillage 3D ellipsoïdal. L'image réelle (en vert) et l'image de synthèse (en rose) sont superposées.

Pour tester l'influence du plaquage de la texture (c'est-à-dire l'alignement de la texture avec le maillage 3D), nous recommençons l'acquisition d'une séquence d'images réelles afin de créer un nouveau modèle 3D avec le maillage précis pour tester une nouvelle fois l'algorithme. La Figure 119 répertorie les estimations avec et sans la méthode *sub-échantillonnage* ainsi que la vérité terrain de cette nouvelle séquence. L'erreur moyenne et l'écart type de l'erreur sont reportés dans le Tableau 10. Enfin, la Figure 120 représente des images réelles de la nouvelle séquence auxquelles nous avons superposé les images de synthèse choisies, aux positions qui minimisent l'écart quadratique pondérée.

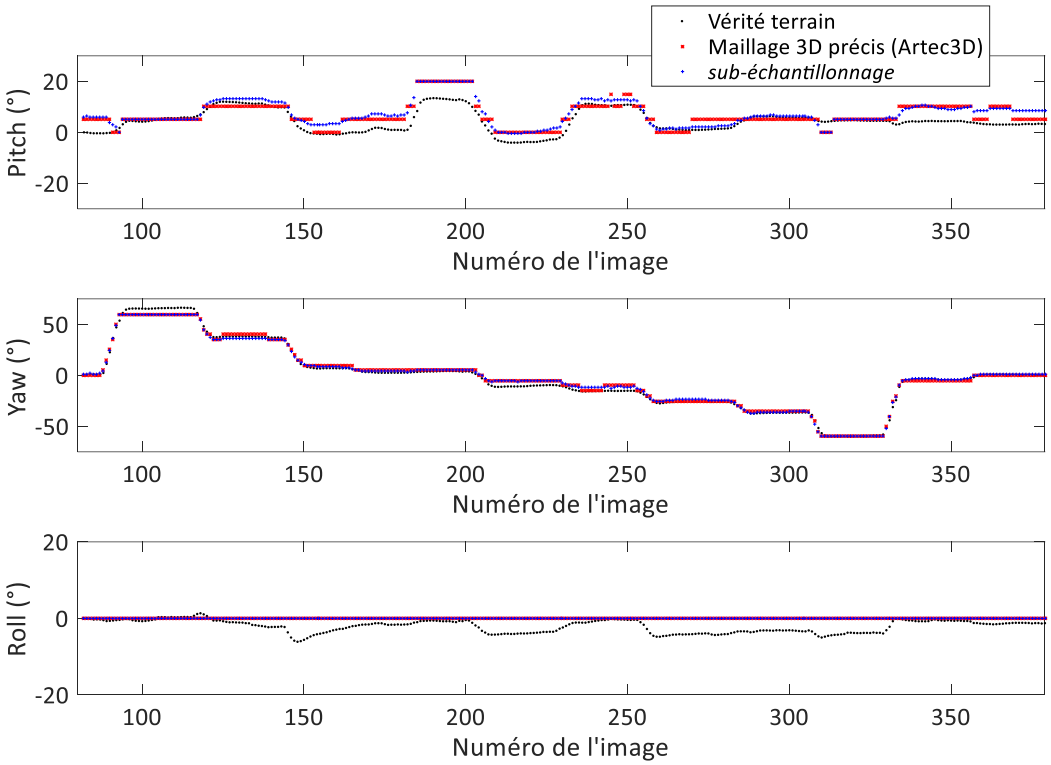


Figure 119. Résultats obtenus avec le maillage 3D précis. Par rapport à la Figure 115, le plaquage de texture a été effectué avec plus de précautions.

Tableau 10. Erreur moyenne et écart type de l'erreur sur l'estimation des angles *yaw* et *pitch* obtenue avec le maillage précis (*Artec3D*). Par rapport au Tableau 8, le plaquage de texture sur le maillage 3D a été réalisé plus précisément.

Méthode *parabolic estimator*	**Résultats des estimations sur les rotations** $(EM \pm S)$		
	Erreur sur *pitch*	**Erreur sur *yaw***	**Erreur sur *roll***
non	-2.1° ± 3,0°	-0.8° ± 2.8°	/
oui	-2.9° ± 2,6°	-0.6° ± 2.6°	/

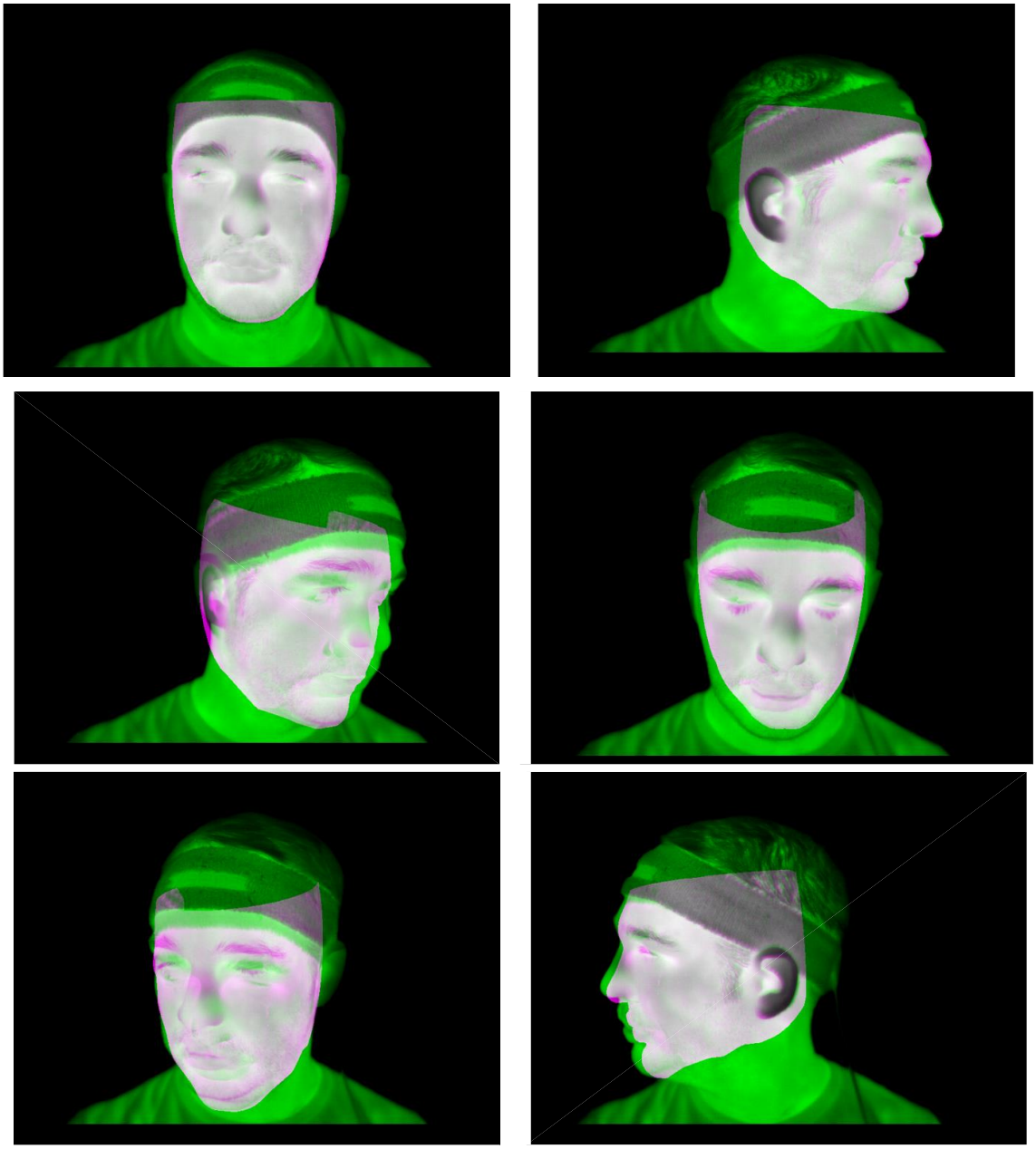


Figure 120. Illustration des résultats obtenus avec le maillage 3D précis. L'image réelle (en vert) et l'image de synthèse (en rose) sont superposées. Par rapport à la Figure 117, le plaquage de texture sur le maillage 3D a été effectué avec plus de précautions.

Ces nouveaux résultats montrent que la méthode d'estimation *parabolic estimator* des angles n'améliore toujours pas la situation de manière claire. Si l'on compare le Tableau 10 au Tableau 8, on constate une amélioration de la moyenne et de l'écart type de l'erreur essentiellement pour l'angle *yaw*. De plus, grâce à la Figure 120, on s'aperçoit que l'estimation du décalage entre l'image réelle et l'image de synthèse est meilleure qu'avec le précèdent modèle thermique (cf. Figure 117).

Enfin pour tester l'effet d'une variation de température entre l'instant de l'acquisition de la texture pour les images de synthèse et l'instant de l'acquisition des images réelles, nous avons placé la caméra dans une chambre climatique à environ 3°C pendant 1h. Puis la caméra a été remontée sur le modèle du véhicule à température ambiante. La température de la caméra augmente tout au long de la séquence vidéo à traiter : elle augmente de $T_{CMin} = 29.2°C$ à $T_{CMax} = 29.9°C$. Les images de synthèses ont été créées à partir d'images acquises à une température de la caméra $T_{synthèse} = 40.8°C$. Il y a donc une différence de température supérieure à 10°C qui provoque inévitablement une variation de la réponse des pixels. Bien évidement l'algorithme basé sur la fonction de coût qui ne prend pas en compte d'offset ni de gain entre les images de synthèse et les images réelles ne fonctionne pas. Nous avons donc testé l'algorithme qui prend en compte un offset (la fonction de coût à minimiser est donc représentée par l'équation (4.15)). Les résultats sont représentés sur la Figure 121 et le Tableau 11.

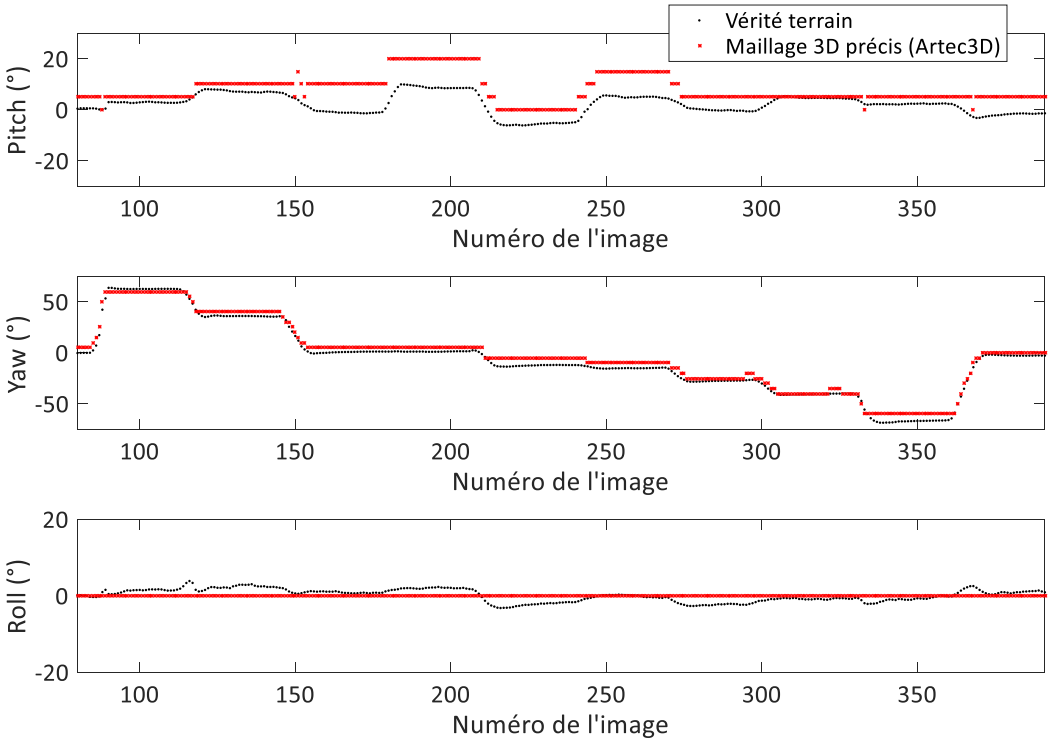


Figure 121. Résultats obtenus avec le maillage 3D précis. Les images de synthèse ont été réalisées avec des images acquises à une température $T_C = 40.8°C$. La température de la caméra lors de l'acquisition des images à traiter varie entre $T_{CMin} = 29.2°C$ à $T_{CMax} = 29.9°C$.

Tableau 11. Erreur moyenne et écart type de l'erreur sur l'estimation des angles *yaw* et *pitch* obtenue avec le maillage précis (*Artec3D*). Les images de synthèse ont été réalisées avec des images acquises à une température $T_C = 40.8°C$. La température de la caméra lors de l'acquisition des images à traiter varie entre $T_{CMin} = 29.2°C$ à $T_{CMax} = 29.9°C$.

Méthode *parabolic estimator*	**Résultats des estimations sur les rotations** $(EM \pm S)$		
	Erreur sur *pitch*	**Erreur sur *yaw***	**Erreur sur *roll***
non	-5.6° ± 3,7°	3.8° ± 2.9°	/

Ces résultats montrent qu'il est possible d'estimer l'orientation du visage par approche « problèmes inverses » avec une caméra non-radiométrique lorsque sa température évolue entre l'instant de l'acquisition de la texture pour la création des images de synthèse et l'instant de l'acquisition des images réelles. L'estimation d'un offset global augmente cependant le temps de calcul. Nous pourrions également prendre en compte un gain pour être plus complet. Là encore le temps de calcul serait augmenté. Le choix de la prise en compte d'un offset (et d'un gain) est à effectuer en considérant les contraintes et les spécifications techniques d'un système final. C'est-à-dire, est-ce qu'il possible d'avoir une caméra radiométrique ? Quelle est la précision de l'orientation du visage nécessaire ? Quelle puissance de calcul est disponible ?

4.4. Conclusion

L'implémentation d'un algorithme basé sur une approche « problèmes inverses » a été présentée dans ce chapitre. Une méthode d'évaluation expérimentale adaptée à la surveillance du conducteur a été mise en place. L'ordre de grandeur de l'erreur sur l'estimation des angles *yaw* et *pitch* est de l'ordre de 3° en moyenne. Celle-ci peut être plus faible si l'étape de plaquage de texture est plus précise. Nous mettons ainsi en évidence l'importance de l'étape de plaquage de texture sur le maillage 3D.

Nous avons considéré dans un premier temps qu'il n'y avait pas d'offset entre l'image réelle et les images de synthèse. Cela implique que si la température de la caméra T_C évolue, cette approche fonctionnera uniquement si un étalonnage radiométrique est disponible (grâce à un élément extérieur de référence, ou grâce à un étalonnage thermique de longue durée).

Comme nous l'avons vu, gérer l'offset entre l'image réelle et les images de synthèse par l'approche « problèmes inverses » est possible mais couteux en temps de calcul. Dans notre application (comme dans bien d'autres), le paramètre d'offset à rechercher est simplement indispensable mais n'est pas utilisé comme information de sortie. Ainsi, dans la référence [134], les auteurs proposent une méthode rapide de normalisation avant de réaliser le produit de corrélation. Cette méthode est connue sous le nom de *normalized crossed-correlation* (*NCC*) dans la littérature anglophone. Elle n'est pas compatible avec un masque de pondération utilisé pour ne pas prendre en compte le fond de l'image. Ainsi, pour qu'elle fonctionne, il est nécessaire de diviser les images en patches pour éviter que le fond ne vienne modifier le coefficient de normalisation des pixels du visage. La *NCC* est employée dans la référence [123] pour estimer

la pose du visage du conducteur en imagerie visible. La *NCC* est une piste potentielle pour répondre à notre problématique. L'idée à retenir est que l'approche par « problèmes inverses » semble pertinente.

L'algorithme de ce chapitre, peut être qualifié de « globale » car il utilise tous les pixels du visage. Une approche d'estimation de la pose qualifiée de « locale » car basée sur des points d'intérêt, sera détaillée au Chapitre 5. Une comparaison entre un algorithme « global » et un algorithme « local » dans le cas de l'imagerie thermique pourra être menée. Les algorithmes « globaux » sont connus pour mieux fonctionner que les algorithmes « locaux » dans le cas d'une faible résolution spatiale [131,135]. Nous avons travaillé avec un format *VGA* (480×640 pixels, $iFOV = 1.4$ mrad). Cette résolution n'est pas considérée comme faible mais le coût de ce type de caméra reste encore élevé. Nous simulerons donc une résolution inférieure au Chapitre 6 et comparerons les algorithmes « globaux » et « locaux ».

Dans notre implémentation, nous testons exhaustivement chacune des 225 images de synthèse de la base. Pour réduire le temps de calcul, nous pourrions adopter une approche pyramidale. Nous pourrions également utiliser la dimension temporelle. Par exemple, dans la littérature scientifique, on trouve des références qui proposent d'utiliser un filtre de Kalman ou un filtre particulaire [123,131]. D'autres travaux proposent de décomposer la base d'images dans différents sous-espaces. Dans les références [136,137] les auteurs utilisent les *eigenvectors*, dans la référence [138] les images sont décomposées grâce aux filtres de *Gabor*.

Chapitre 5. Implémentation du suivi de la pose du visage par appariement de points clefs 3D-2D

5.1. Introduction

L'objectif de ce chapitre est de présenter la méthode d'estimation de l'orientation et de la position du visage en se basant sur la mise en correspondance de points d'intérêts. Lorsque que nous ferons référence à cette méthode dans la suite du manuscrit, nous utiliserons la formulation *appariement de points 3D-2D*. Nous assumons disposer d'un maillage 3D texturé et d'une base d'images synthétiques du visage, différemment orienté en *pitch* et en *yaw,* tel que cela a été expliqué au chapitre 3. La particularité de la méthode d'*appariement de points 3D-2D* est qu'elle s'appuie sur l'analyse locale de l'image. Nous débuterons ce chapitre par un rappel sur les principales méthodes utilisées dans le domaine de la vision par ordinateur pour détecter et mettre en correspondance des points d'intérêt (ou points clefs). Puis nous expliquerons, en nous appuyant sur des tests, pourquoi nous avons choisi d'utiliser la méthode *SIFT*. Les points clefs $\boldsymbol{q}_i^v$ détectés sur les images synthétiques peuvent être exprimés en 3D grâce au maillage 3D qui a permis la création des images (cf. Figure 122). Les points clefs $\boldsymbol{q}_i^c$ détectés sur l'image questionnée sont, quant à eux, exprimés dans le repère 2D du plan image de la caméra. Notre objectif est d'estimer la matrice de rotation $\boldsymbol{R}$ ainsi que le vecteur de translation $\boldsymbol{t}$ entre le repère de la caméra et celui de la tête à partir des appariements de points 3D-2D. Nous détaillerons en particulier dans la section 5.3 une méthode de l'état de l'art qui permet d'estimer la pose dans un contexte de mise en correspondance de points d'intérêt difficile (ce contexte est propre à l'utilisation de caméras thermiques non-refroidies pour imager un visage humain). Enfin, nous détaillerons dans la section 5.4 les spécificités de l'implémentation, liées à notre application. Nous discuterons également de la possibilité de supprimer des *outliers* dans certaines conditions en imagerie thermique en nous appuyant sur les valeurs des pixels (compensation de l'offset globale en

fonction de la température de la caméra). Puis, dans la section 5.5, nous comparerons les estimations de l'orientation aux mesures de la centrale inertielle.

5.2. Les détecteurs et les descripteurs de points d'intérêt

Un point d'intérêt (on trouve également les termes *local feature* ou *interest point* dans la littérature anglophone) désigne une 'petite' zone de l'image qui diffère de son entourage direct. Dans notre cas il s'agira d'une variation de niveau de gris. Un point d'intérêt peut être une tâche, mais aussi un bord, ou une zone de texture particulière. Le détecteur localise la région de l'image intéressante. Ensuite le descripteur la rend identifiable. Pour notre application, il est intéressant de pouvoir détecter et décrire les points d'intérêts afin de les utiliser comme des points d'ancrage pour analyser le mouvement du visage du conducteur. Ce qui nous intéresse, ce n'est pas ce que représentent les points, mais c'est le fait qu'ils soient

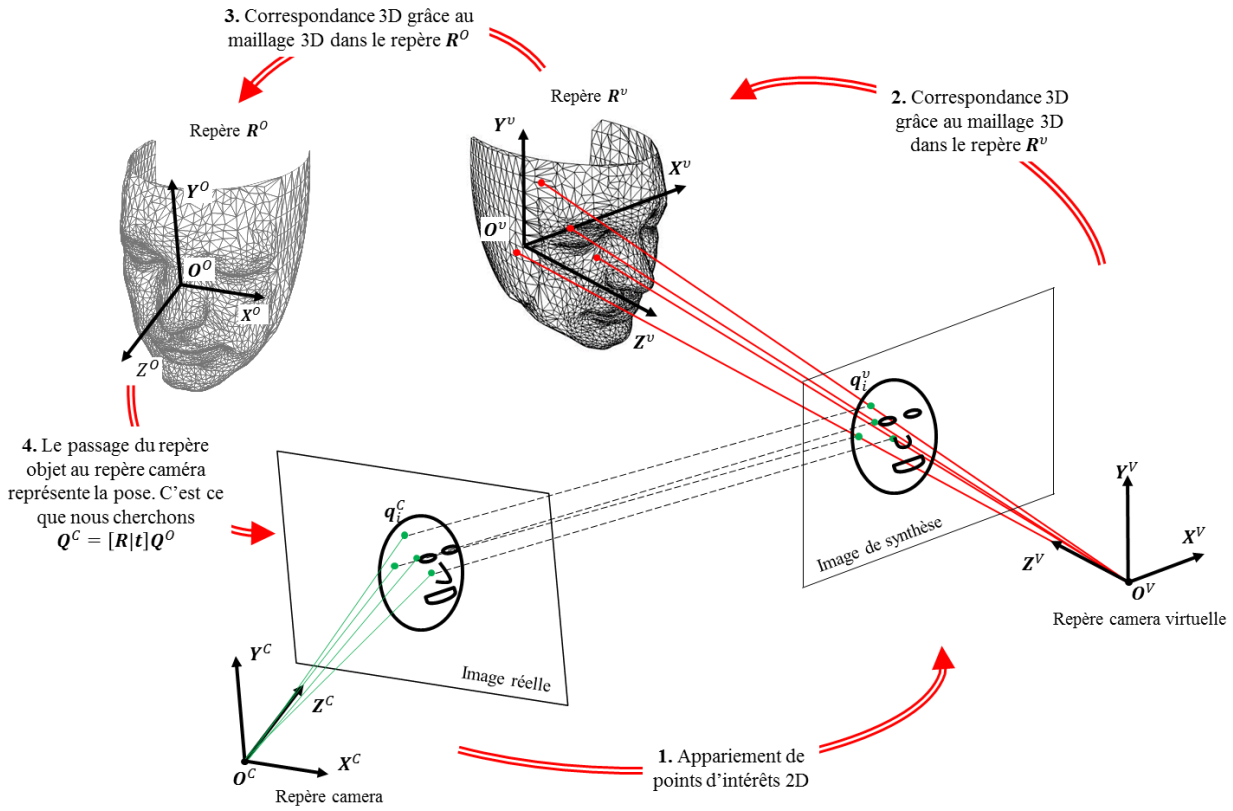


Figure 122. Principe général de l'utilisation d'un maillage 3D pour estimer la pose $[\boldsymbol{R}|\boldsymbol{t}]$ liant le repère caméra au repère objet.

identifiables. Ainsi, lorsqu'un même point est détecté sur deux images prises dans des conditions différentes (point de vue, illumination lorsqu'une caméra visible est utilisée, bruit…), le descripteur permet d'assurer qu'il s'agit toujours du même point. Les points d'intérêts sont couramment utilisés dans les applications d'estimation de la pose.

Dans les références [139,140] des analyses comparatives de couple de détecteur/descripteur pour l'imagerie thermique ont été menées. Ces travaux montrent que les algorithmes de détection/description *SIFT*, ainsi que sa version plus rapide *SURF*, semblent les mieux adaptés à l'imagerie thermique.

Dans cette section, notre objectif est de rappeler les méthodes de détection et de description qui permettent d'apparier (ou de mettre en correspondance) des points d'intérêts dans deux images. Le choix de la combinaison du détecteur et du descripteur est réalisé en fonction de la nature de l'image à traiter (bande spectrale, bruit, objet à tracker…) et des invariances nécessaires pour l'application. Nous verrons dans les détails de l'implémentation de notre méthode (cf. section 5.4) qu'il est nécessaire d'être invariant à des changements de points de vue. Dans notre cas, une invariance à +/- 5° suffit. Une invariance plus importante améliore le temps de calcul car on pourrait réduire la taille de la base d'images de synthèses. De plus, elle améliore la robustesse de l'algorithme final d'estimation de la pose. Nous démontrerons par des tests pourquoi nous avons choisi d'utiliser l'algorithme de détection/description *SIFT*.

5.2.1. Les détecteurs de points d'intérêt

Il existe de nombreuses méthodes pour détecter des points d'intérêts. Un état de l'art datant de 2008 est dressé dans [141]. Les auteurs proposent une classification en fonction de la nature de la zone locale détectée : coin, tâche (on trouve dans littérature anglophone le terme *blob*) ou région. Nous allons nous intéresser uniquement à quatre détecteurs connus : *Harris*, *Harris-Laplace*, *LoG* et *DoG*. Le détecteur de *Harris* est un détecteur de coin. La méthode *Harris-Laplace* est un détecteur multi-échelle de coins. La méthode laplacien de gaussienne *(LoG* en abrégé dans la littérature anglophone*)* est un détecteur multi-échelle de tâche. La méthode appelée différence de gaussiennes (méthode *DoG* en abrégée dans la littérature anglophone), est également un détecteur multi-échelle de tâches. Le détecteur *DoG* est utilisé dans l'algorithme appelé *scale-invariant feature transform* (*SIFT* en abrégé dans la littérature anglophone) qui est une méthode complète de détection et de description de points d'intérêt. Le *DoG* est une manière efficace en temps de calcul d'approximer le *LoG*. Le choix du détecteur est important car il impacte les performances du descripteur et finalement le nombre de points mis en correspondance avec succès entre deux images.

5.2.1.1. Le détecteur de Harris

Dans [142] Harris reprend et améliore les travaux de Moravec [143]. Il développe un détecteur permettant d'extraire des coins et/ou les bords d'une image et améliore la robustesse au bruit. Son détecteur est basé sur les résultats des produits de convolution entre l'image et les dérivées premières de la fonction Gaussienne long des axes *x* et *y*. La Gaussienne permet de réduire les effets néfastes du bruit. On note L la convolution de l'image avec une gaussienne d'échelle σ notée G :

$$L(x, y) = G(x, y, \sigma) * I(x, y),$$

avec

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

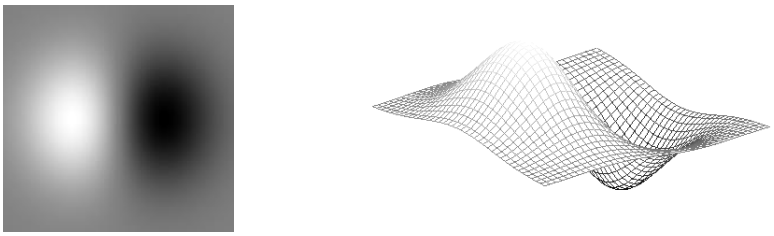


Figure 123. Dérivée première d'une gaussienne selon l'axe horizontal. Elle est notée g_x.

Les dérivées premières de la fonction gaussienne selon les axes x et y sont respectivement notées g_x et g_y. Nous rappellerons simplement l'expression de g_x :

$$g_x(x, y, \sigma) = \frac{\partial G}{\partial x} = \frac{-x}{2\pi\sigma^4} e^{-(x^2+y^2)/\,2\sigma^2}$$

On note L_x la convolution de l'image avec la dérivée de la gaussienne d'échelle σ notée g_x :

$$L_x(x, y, \sigma) = g_x(x, y, \sigma) * I(x, y)$$

On remarquera, comme illustré sur la Figure 123, que g_x n'est pas isotrope (il en est de même pour g_y).

Le détecteur de *Harris* est basé sur la matrice des moments d'ordre 2 qui décrit la distribution locale du gradient, lissée par une Gaussienne d'échelle σ :

$$\mu(x, y) = \begin{bmatrix} L_x^2(x, y) & L_x L_y(x, y) \\ L_x L_y(x, y) & L_y^2(x, y) \end{bmatrix}, \tag{5.1}$$

La réponse du filtre de *Harris* est décrite par l'équation ci-dessous :

$$R_{Harris}(x, y) = det(\mu(x, y)) - \alpha\, trace^2(\mu(x, y)) \tag{5.2}$$

La réponse R_{Harris} est positive dans les régions des coins et négative dans les régions des bords. Le paramètre α permet de contrôler la sensibilité de la détection de coins : lorsqu'il est petit, les coins sont détectés plus facilement. Lorsque les points d'intérêts d'une image ressemblent plus à des tâches qu'à des coins ou des bords, on utilise un détecteur de tâche comme les détecteurs *LoG* ou *DoG*.

5.2.1.2. Le détecteur *Laplacian of Gaussian (LoG)*

Afin d'extraire des points clefs d'une image, les dérivées secondes sont utilisées depuis déjà un certain temps [144, 145]. Les dérivées secondes de la fonction gaussienne sont notées comme suit :

$$g_{xx} = \frac{\partial^2 G}{\partial x^2},$$

$$g_{yy} = \frac{\partial^2 G}{\partial y^2},$$

$$g_{xy} = \frac{\partial^2 G}{\partial x \partial y}.$$

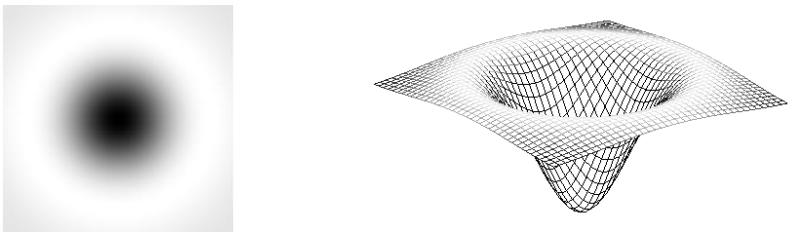


Figure 124. Laplacien d'une gaussienne. Ce noyau est utilisé pour détecter les points d'intérêt de l'image.

On note L_{xx}, L_{yy} et L_{xy} les convolutions de l'image avec les dérivées secondes de la fonction gaussienne d'échelle σ :

$$\begin{aligned} L_{xx}(x,y,\sigma) &= g_{xx}(x,y,\sigma) * I(x,y), \\ L_{yy}(x,y,\sigma) &= g_{yy}(x,y,\sigma) * I(x,y), \\ L_{xy}(x,y,\sigma) &= g_{xy}(x,y,\sigma) * I(x,y). \end{aligned} \tag{5.3}$$

Introduisons la matrice hessienne dont les composantes peuvent être utilisées pour détecter des structures de l'image contenant de l'information grâce aux dérivées secondes de la gaussienne :

$$\mathcal{H}(x,y,\sigma) = \sigma^2 \begin{bmatrix} L_{xx}(x,y,\sigma) & L_{xy}(x,y,\sigma) \\ L_{xy}(x,y,\sigma) & L_{yy}(x,y,\sigma) \end{bmatrix} \tag{5.4}$$

Notons que dans l'équation (5.4) un facteur σ^2 apparaît. Il est lié à la notion d'espace d'échelle que nous allons introduire ici.

La notion d'espace d'échelle est formulée par Witkin dans [146]. C'est l'idée que l'information utile d'une image peut être située à plusieurs échelles qu'il est impossible de prédire à l'avance. En d'autres termes, un point d'intérêt peut être « petit » ou « gros » et nous ne le savons pas par avance. Witkin préconise donc d'explorer cette dimension en appliquant des filtres passes bas dont la taille du noyau est variable.

Koenderink, Lindeberg ou encore Florack [147, 148, 149] ont montré que le filtrage de l'image grâce à un noyau Gaussien est une manière d'explorer l'espace d'échelle d'une image. σ est donc le paramètre qui contrôle l'échelle du noyau Gaussien.

Afin de rendre la détection invariante à l'échelle, la normalisation a été ajoutée dans l'équation (5.4) grâce au facteur σ^2. Lindeberg détecte les tâches d'une image grâce au laplacien d'une gaussienne normalisé par rapport à l'échelle, qui n'est autre que la trace de la matrice hessienne [150]. Cette technique est appelée *LoG* (*Laplacian of Gaussian*). La réponse d'un filtre *LoG* s'exprime comme suit :

$$R_{LoG}(x,y) = \sigma^2 \left| L_{xx}(x,y,\sigma) + L_{yy}(x,y,\sigma) \right| \tag{5.5}$$

Le noyau laplacien possède une taille caractéristique et une symétrie circulaire (cf. Figure 124). Lorsque la taille et la structure d'une zone de l'image correspond au noyau, la grandeur (5.5) atteint un extremum. C'est pourquoi le *LoG* est bien adapté à la détection de tâche. Cependant comme le montre Mikolajczyk [151], les coins, les bords et les jonctions multiples peuvent également être détectés.

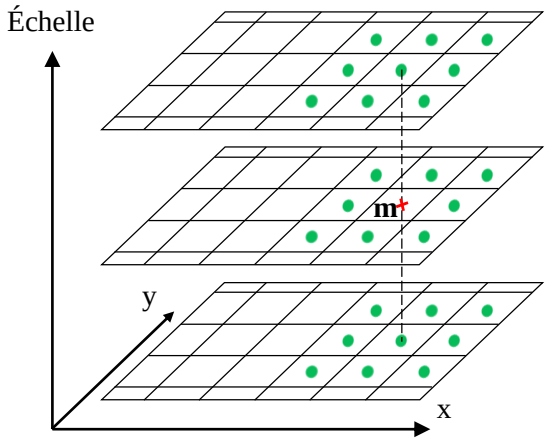


Figure 125. Détection d'un point d'intérêt dans un espace 3D (x, y, σ).

Lindeberg [150] calcule R_{LoG} pour un ensemble d'échelles et cherche de manière automatique les points d'intérêts (cf. Figure 125) : un point de coordonnées (x_m, y_m, σ_n) est détecté si

$$R_{LoG}(x_m, y_m, \sigma_n) > R_{LoG}(x_i, y_j, \sigma_l) \quad \forall \quad i,j \in \{m-1, m, m+1\} \quad et \quad l \in \{n-1, n, n+1\},$$

et si

$$D(x_m, y_m, \sigma_n) > seuil$$

Mikolajczyk montre dans ses travaux de thèse [151] que la recherche automatique de l'échelle est meilleure en utilisant la réponse du filtre *LoG*, R_{LoG} qu'en utilisant d'autres filtres basés sur les dérivées premières d'une gaussienne telle que le filtre de *Harris*.

5.2.1.3. Le détecteur *Difference-of-Gaussians (DoG)* implémenté dans *SIFT*

Le détecteur *DoG* proposé par Lowe [152] est une implémentation efficace du *LoG*. En effet, l'intérêt du *DoG* est qu'il permet d'approximer le laplacien d'une gaussienne en un temps de calcul contenu.

Nous allons détailler le *DoG*. Il s'appuie sur le fait que la dérivée de la gaussienne par rapport à l'échelle est égale au laplacien d'une gaussienne multipliée par l'échelle :

$$\frac{\partial G}{\partial \sigma} = \sigma(g_{xx} + g_{yy}) \tag{5.6}$$

Exprimons les différents membres de l'équation (5.6) pour visualiser plus facilement cette égalité :

$$g_{xx} = \frac{\partial^2 G}{\partial x^2} = \frac{1}{2\pi\sigma^4} e^{-(x^2+y^2)/2\sigma^2} \left(\frac{x^2 - \sigma^2}{\sigma^2}\right).$$

Ainsi,

$$g_{yy} + g_{xx} = \frac{1}{2\pi\sigma^4} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \left(\frac{x^2 + y^2 - 2\sigma^2}{\sigma^2}\right)$$

Et comme,

$$\frac{\partial G}{\partial \sigma} = \frac{1}{2\pi\sigma^3} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \left(\frac{x^2 + y^2 - 2\sigma^2}{\sigma^2} \right),$$

La relation est vérifiée.

En pratique la dérivée est approximée par une différence finie :

$$\frac{\partial G}{\partial \sigma} \approx \frac{G(x, y, k_{i+1}\sigma) - G(x, y, k_i\sigma)}{k_{i+1}\sigma - k_i\sigma}. \tag{5.7}$$

Les facteurs $k_i = 2^{i/s}$ et $k_{i+1} = 2^{(i+1)/s}$ sont définis pour $i\epsilon[1, \dots, s-1]$. Finalement, en combinant les équations (5.6) et (5.7), on remarque que la différence de gaussienne est proportionnelle au laplacien normalisé :

$$G(x, y, k_{i+1}\sigma) - G(x, y, k_i\sigma) \approx (k_{i+1} - k_i)\sigma^2 \nabla^2 G(x, y, k_i\sigma). \tag{5.8}$$

Pour explorer l'espace des échelles, l'image initiale va être convoluée à une série de fonctions gaussiennes d'échelles variables (colonne de gauche de la Figure 126). Les images adjacentes séparées d'un facteur d'échelle $k = 2^{1/s}$ sont soustraite entres elles pour approximer le laplacien normalisé (colonne de droite de la Figure 126). Le paramètre s permet de modifier le nombre d'échelles explorées dans un *octave*. Un *octave* est un ensemble d'échelles comprises entre une valeur donnée σ_0 et son double $2\sigma_0$. Le nombre d'octaves et le paramètre s permettent de contrôler le nombre total d'échelles explorées. Ils sont donc à fixer par l'utilisateur en fonction de son application. Concernant le paramètre s, dans la plupart des cas, Lowe montre qu'il n'est pas nécessaire de l'augmenter au-delà de 3. En effet, augmenter s permet détecter plus de points, mais leur robustesse est moins bonne. Pour appuyer ses propos, il utilise le critère de répétabilité (nous l'utilisons également pour choisir un détecteur parmi ceux présentés dans ce manuscrit dans la section 5.2.1.5).

Finalement, pour approximer le laplacien, le produit de convolution entre la différence de gaussienne et l'image initiale, notée D, est calculé :

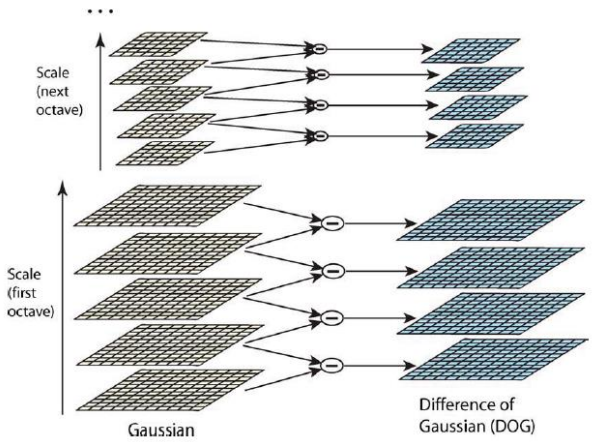


Figure 126. Approximation du Laplacien par la différence de Gaussienne.

$$D(x, y, k_i\sigma) = \big(G(x, y, k_{i+1}\sigma) - G(x, y, k_i\sigma)\big) * I(x, y),$$

$$D(x, y, k_i\sigma) = L(x, y, k_{i+1}\sigma) - L(x, y, k_i\sigma) \quad (5.9)$$

Comme Lindeberg, Lowe propose de rechercher automatiquement les points d'intérêts dans l'espace 3D (x, y, σ). Dans la Figure 125, un point est détecté si $D(x_m, y_m, \sigma_n) > D\big(x_i, y_j, \sigma_l\big)$ avec $i, j \in \{m-1, m, m+1\}$ et $n \in \{l-1, l, l+1\}$ et si $D(x_m, y_m, \sigma_n) > seuil$.

Les points d'intérêt situés sur les arrêtes sont souvent considérés comme instables car leurs localisations précises sont difficiles. En effet, les réponses des détecteurs de points d'intérêts le long d'une arrête sont proches. Une densité importante de points d'intérêt est donc situé le long des arrêtes. De plus, les descripteurs sont incapables de différencier de manière robuste les points situés le long des arrêtes. Une étape supplémentaire de suppression des points détectés sur les arrêtes est ajoutée dans *SIFT*. A la manière du détecteur de Harris, le rapport des courbures principales est calculé, grâce aux composantes de la matrice hessienne. Ce rapport est élevé pour les points situés sur les arrêtes. Un seuillage permet donc de les supprimer.

Enfin, pour augmenter la précision de la localisation des points d'intérêt, une interpolation grâce à un développement en série de Taylor à l'ordre 2 est utilisée. Pour un point d'intérêt donné qui prend une valeur D_a aux coordonnées $\boldsymbol{x}_a = [x_a, y_a, \sigma_a]^T$ et pour lequel on note $\boldsymbol{x} = [x - x_a, y - y_a, \sigma - \sigma_a]^T$ un décalage en coordonnées spatiales et en échelle par rapport à $\boldsymbol{x}_a$, le développement en série de Taylor s'exprime comme suit :

$$D_a(\boldsymbol{x}) = D_a + \frac{\partial D_a^T}{\partial \boldsymbol{x}}\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T\frac{\partial^2 D_a}{\partial \boldsymbol{x}^2}\boldsymbol{x}. \quad (5.10)$$

La dérivée de cette expression est nulle lorsque $D_a(\boldsymbol{x})$ atteint un extremum autour du point $\boldsymbol{x}_a$. Ainsi, l'estimation de la localisation de cet extremum noté $\widehat{\boldsymbol{x}}$ s'exprime comme suit :

$$\widehat{\boldsymbol{x}} = -\left(\frac{\partial^2 D_a}{\partial \boldsymbol{x}^2}\right)^{-1}\frac{\partial D_a}{\partial \boldsymbol{x}} \quad (5.11)$$

5.2.1.4. Le détecteur Harris-Laplace

Mikolajczyk [151] propose de combiner le détecteur de Harris avec une recherche dans l'espace des échelles grâce à l'opérateur laplacien. Mikolajczyk justifie le choix de cette combinaison par la supériorité du détecteur de Harris dans l'espace spatial. Cependant, dans l'espace des échelles, le détecteur de Harris n'atteint pas souvent un maximum, ce qui rend la sélection des points d'intérêt difficile. Au contraire, il identifie le détecteur *LoG* comme étant le plus efficace pour la recherche automatique de l'échelle.

5.2.1.5. Choix du détecteur

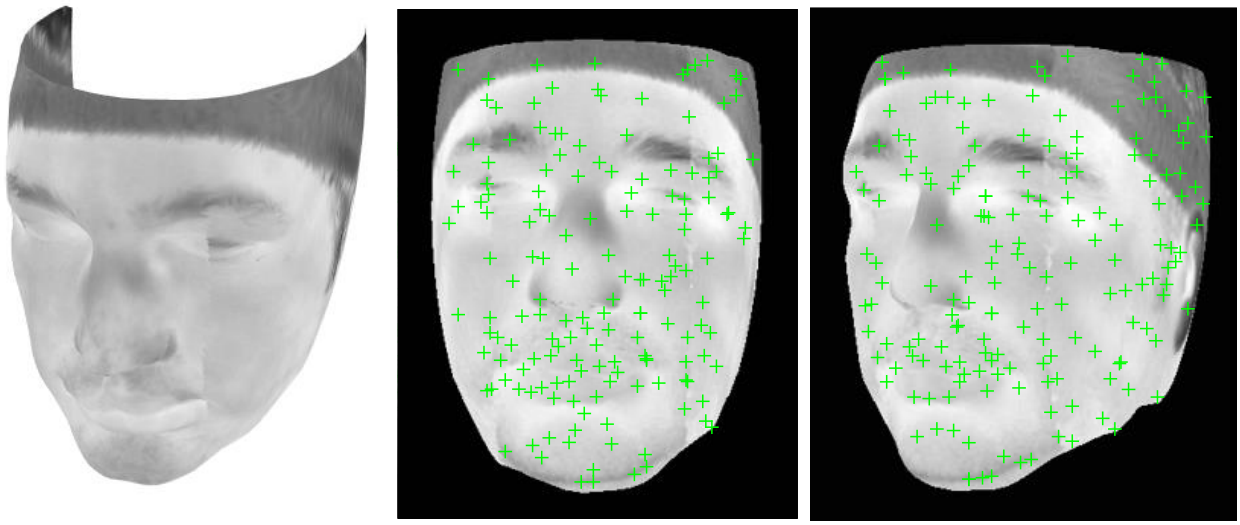
Nous souhaitons comparer dans cette section les détecteurs de *Harris*, *Harris-Laplace* et *DoG*. Pour *Harris*, nous utilisons le code fourni par *Matworks* dans MATLAB (fonction *corner*). Nous utilisons pour *Harris-Laplace* des codes inspirés de l'implémentation Matlab de Vincent Garicia [153]. Pour le détecteur *DoG*, nous utilisons le code proposé dans la *toolbox VLFEAT* [154]. Nous avons identifié que la plupart des points d'intérêt se trouvent à l'échelle $\sigma = 3$. Le détecteur *Harris* est utilisé à cette échelle. L'espace des échelles du détecteur *Harris-Laplace* est utilisé pour une seule octave qui débute à $\sigma_0 = 2$. On utilise 4 sous-échelles : $\sigma \epsilon$ [2 , 2.52 , 3.18 , 4].

Le détecteur *DoG* est utilisé avec les paramètres par défaut implémentés dans la *toobox VLFEAT* (? octaves, 3 sous-niveaux).

Pour comparer les performances des détecteurs, l'indicateur proposé dans [155] est utilisé. Il s'agit du critère de répétabilité. Il va compter le nombre de points **correspondants** détectés sur deux images d'un même objet. Deux points 2D détectés sur deux images prises de deux points de vue différents sont définis comme **correspondants** s'ils appartiennent au même point 3D de l'objet observé.

Remarque : le descritpeur de points d'intérêt, décrit dans la section suivante, n'entre pas en considération dans l'analyse de la performance du détecteur de points d'intérêt. En effet, toute l'astuce de la méthode que nous allons décrire, est d'être capable de mettre en correspondance des points d'intérêt sans descripteur.

Nous disposons d'un maillage 3D texturé. Nous proposons donc de créer deux images, respectivement $\boldsymbol{I}_1$ et $\boldsymbol{I}_2$, en projetant le maillage 3D grâce aux matrices de projections, respectivement $\boldsymbol{P}_1$ et $\boldsymbol{P}_2$ (cf. Figure 127). Nous utilisons les notations définies sur la Figure 128. Un point noté $\boldsymbol{x}_i^{(1)}$ est détecté sur l'image $\boldsymbol{I}_1$ et un point noté $\boldsymbol{x}_j^{(2)}$ est détecté sur l'image $\boldsymbol{I}_2$. Le point noté $\boldsymbol{x}_i^{(1p2)}$ est un point détecté sur l'image $\boldsymbol{I}_1$, puis projeté sur l'image $\boldsymbol{I}_2$. Finalement la distance $\overline{\boldsymbol{x}_i^{(1p2)}\boldsymbol{x}_j^{(2)}}$ va nous permettre de déterminer si les points $\boldsymbol{x}_i^{(1)}$ et $\boldsymbol{x}_j^{(2)}$ sont des correspondants, c'est-à-dire $\overline{\boldsymbol{x}_i^{(1p2)}\boldsymbol{x}_j^{(2)}} < \epsilon$.



a) b) c)

Figure 127. Méthode pour créer les images utilisées pour établir le critère de répétabilité. La colonne a) représente le maillage 3D. La colonne b) représente la projection du maillage 3D avec un angle *yaw* qui vaut 0° (on peut la définir comme l'image notée I_1). La colonne c) représente la projection du maillage 3D avec un angle yaw qui vaut -22.5° (on peut la définir comme l'image notée I_2). Les croix vertes représentent les points d'intérêt détectés grâce à l'opérateur *DoG*.

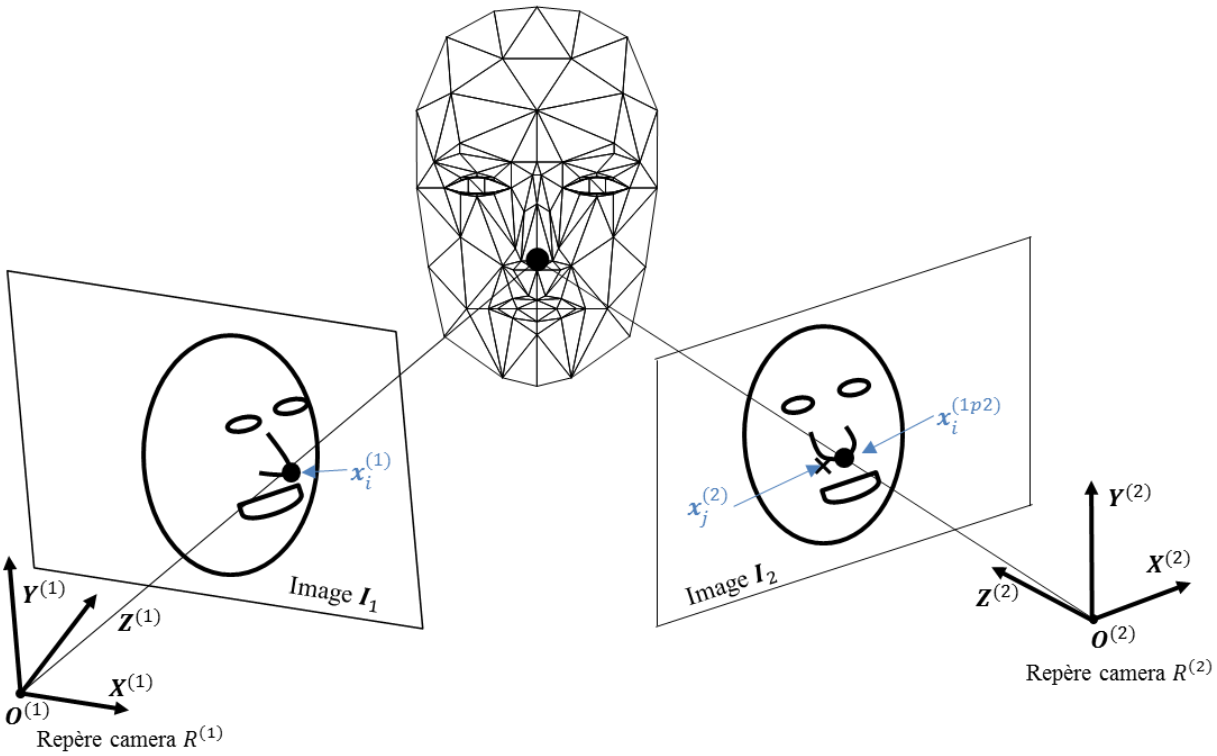


Figure 128. Principe de mesure de la répétabilité d'un point d'intérêt détecté sur deux images d'un même objet acquises de points de vue différents.

Pour récapituler, établir la correspondance entre des points détectés sur deux images peut être réalisé comme suit :

a. détecter des points 2D $\boldsymbol{x_i^{(1)}}$ sur l'image $\boldsymbol{I_1}$,
b. approximer leurs coordonnées 3D notées $\boldsymbol{Q_i^O}$,
c. projeter les points 3D $\boldsymbol{Q_i^O}$ sur l'image $\boldsymbol{I_2}$ grâce à la matrice de projection $\boldsymbol{P_2}$:

$$\begin{bmatrix} \boldsymbol{x_i^{(1p2)}} \\ 1 \end{bmatrix} = \boldsymbol{K_{Gobi}}.\boldsymbol{P_2}.\boldsymbol{Q_i^O},$$

d. détecter des points 2D $\boldsymbol{x_j^{(2)}}$ sur l'image $\boldsymbol{I_2}$,
e. calculer les distances $\overline{\boldsymbol{x_i^{(1p2)}}\boldsymbol{x_j^{(2)}}}$ en pixel,
f. Compter le nombre n_d de distances qui vérifient $\overline{\boldsymbol{x_i^{(1p2)}}\boldsymbol{x_j^{(2)}}} < \varepsilon$.

Apportons quelques précisions sur l'étape b. Compte tenu du caractère discret du maillage 3D, nous ne pouvons qu'approximer les coordonnées 3D $\boldsymbol{Q_i^O}$ correspondant au point 2D $\boldsymbol{x_i^{(1)}}$ détectés sur l'images $\boldsymbol{I_1}$. Cette approximation consiste à choisir parmi les *vertices* du maillage 3D $\boldsymbol{Q_k^O}$ celui dont la projection sur l'image $\boldsymbol{I_1}$, notée $\boldsymbol{q_k^{(1)}}$, est la plus proche de $\boldsymbol{x_i^{(1)}}$. En pratique 70 000 *vertices* composent le modèle 3D (ce sont les $\boldsymbol{Q_k}$). Nous venons d'introduire la notation $\boldsymbol{u_k^{(1)}}$ qui représente la projection du *vertice* $\boldsymbol{Q_k^O}$ du maillage 3D sur l'image $\boldsymbol{I_1}$:

$$\boldsymbol{q_k^{(1)}} = \boldsymbol{P_1}\boldsymbol{Q_k^O} \tag{5.12}$$

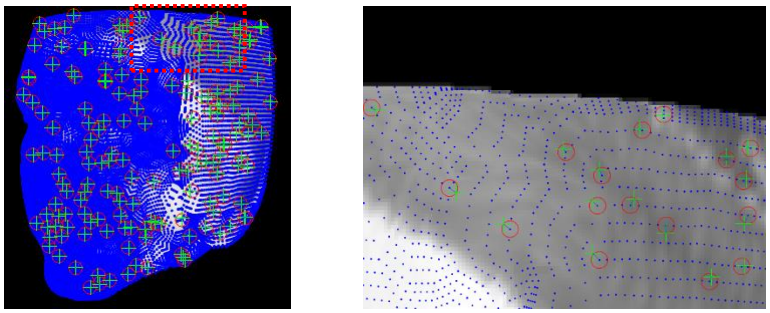


Figure 129. Recherche des coordonnées 3D d'un point d'intérêt détecté sur une image de synthèse 2D issue du modèle 3D. Les points bleus sont les projections $\boldsymbol{q}_k$ des points du maillage 3D $\boldsymbol{Q}_k$ du modèle 3D. Les croix vertes sont les points d'intérêt $\boldsymbol{x}_i$ détectés sur l'image. Les cercles rouges entourent les points du modèle 3D qui approximent les points détectés sur l'image 2D.

Pour choisir parmi les $\boldsymbol{Q}_k^O$, on étudie l'ensemble des 70 000 distances s_k et on relève l'indice *k* qui minimise cette distance :

$$\text{On cherche } k \text{ tel que } s_k = \min_{k\epsilon[1,70\,000]} \left(s_k = \left|\overrightarrow{\boldsymbol{x}_i^{(1)}\boldsymbol{q}_k^{(1)}}\right|\right) \tag{5.13}$$

Finalement le point 3D qui approxime au mieux le point 2D détecté $\boldsymbol{x}_i^{(1)}$ est celui qui minimise s_k (cf. Figure 129). L'erreur introduite par cette approximation est de 0.87 ± 0.60 pixels (moyenne ± écart type) sur l'ensemble des points de l'image en Figure 129.

Le taux de répétabilité est défini comme étant le rapport du nombre de points n_d détectés dans les deux images à une erreur de localisation près ε, sur le nombre de points détectés au total et visibles sur les deux images :

$$r = \frac{n_d(\varepsilon)}{min(n_1, n_2)} \tag{5.14}$$

Le nombre de points détectés sur l'image $\boldsymbol{I}_1$ visibles sur l'image $\boldsymbol{I}_2$ est noté n_1. De même, le nombre de points détectés sur l'image $\boldsymbol{I}_2$ visibles sur l'image $\boldsymbol{I}_1$ est noté n_2. Pour déterminer si un point détecté sur une image est visible sur une autre image, on évalue l'angle entre l'axe optique de la *caméra virtuelle* et la normale de la *face*. Si cet angle est inférieur à 90° avec l'axe optique alors le point est considéré comme visible.

Le taux de répétabilité est calculé entre deux images $\boldsymbol{I}_1$ et $\boldsymbol{I}_2$ dont l'orientation relative du visage évolue. Nous menons deux expériences :

- **Dans la première**, l'image $\boldsymbol{I}_1$ est un visage de synthèse dont l'orientation est nulle en *yaw* et *pitch* et l'image $\boldsymbol{I}_2$ est un visage de synthèse dont l'orientation selon l'angle *yaw* évolue de -30° à +30° par pas de 2.5° (l'angle *pitch* reste nul). Deux points sont considérés répétés si $\left|\overrightarrow{\boldsymbol{x}_i^{(1p2)}\boldsymbol{x}_i^{(2)}}\right| < \varepsilon$ avec ε = 2.5 pixels. Nous introduisons sur chaque image un bruit blanc gaussien d'écart type 10 Adu (sur 16 bits) pour modéliser la *NETD*. Le taux de répétabilité pour les différents détecteurs est illustré sur la Figure 130 a). Le nombre de points total répétés est illustré sur Figure 130 b).

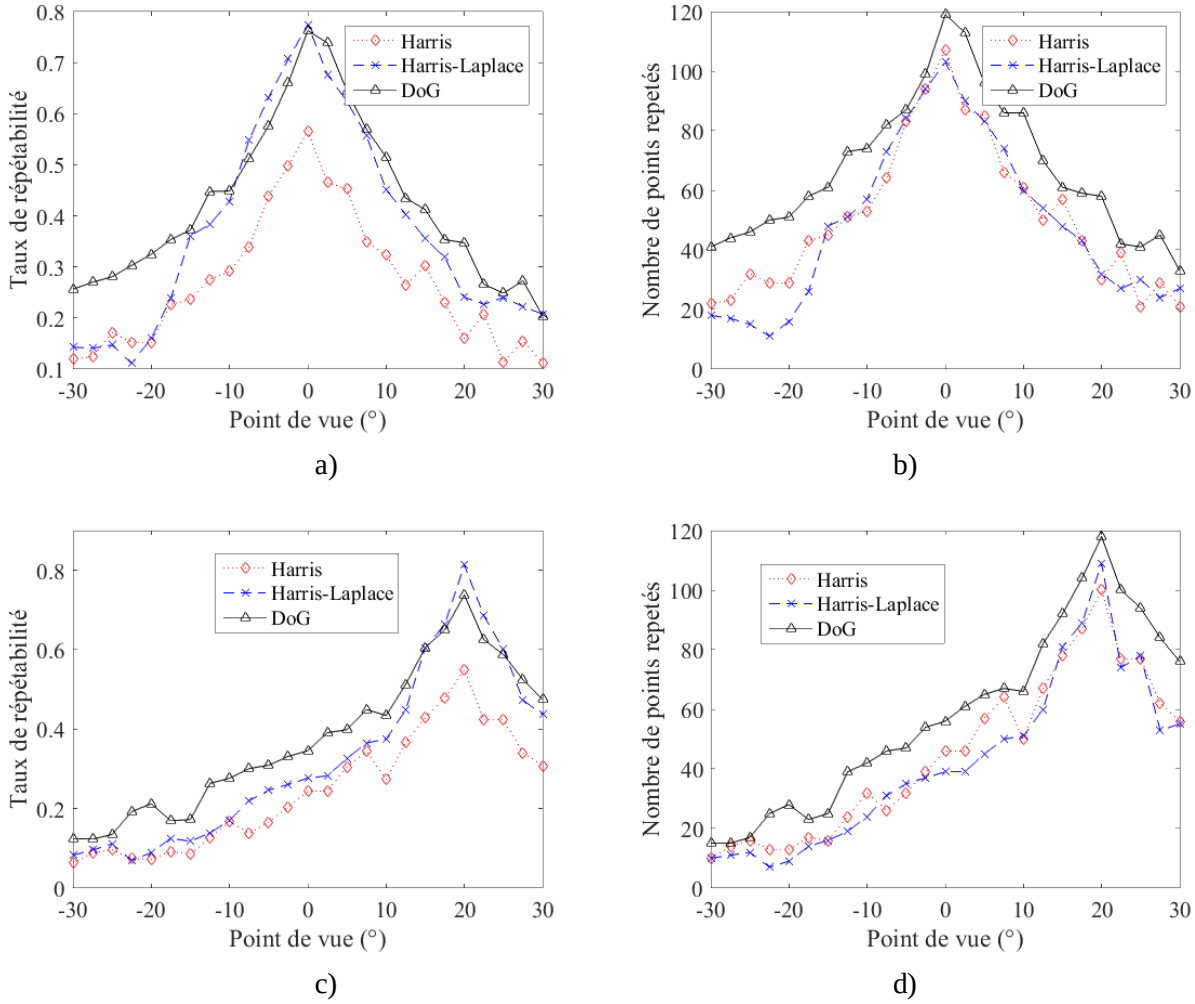


Figure 130. Evaluation de la répétabilité pour plusieurs types de détecteurs : Harris, Harris-Laplace et *DoG*. Les figures a) et c) présentent le taux de répétabilité. Les figures b) et d) présentent le nombre total de points répétés. Sur a) et b), on évalue la répétabilité entre une image de face du visage et des images dont l'orientation du visage en yaw varie de -30 à +30° par pas de 2.5°. Sur c) et d), on évalue la répétabilité entre une image du visage orientée à 20° en yaw et des images dont l'orientation du visage en yaw varie de -30 à +30° par pas de 2.5°.

- **La seconde expérience** est équivalente à la première mis à part que dans l'image $\boldsymbol{I}_1$, l'orientation du visage vaut 20° en *yaw* et 0° en *pitch*.

La Figure 130 a) et c) montre que le détecteur *DoG* possède le deuxième meilleur taux de répétabilité lorsque le point de vue varie entre +/-10°. Au-delà, de +/-10° le détecteur *DoG* possède le meilleur taux de répétabilité. La Figure 130 b) et d) montre que la méthode *DoG* détecte le plus grand nombre de points répétables quelque-soit la variation de point de vue.

Les méthodes de *Harris-Laplace* et *Harris* détectent approximativement le même nombre de points répétables (Figure 130 a) et c)). Le détecteur de *Harris-Laplace* possède cependant un meilleur taux de

répétabilité. Ceci nous permet de dire que la recherche dans l'espace des échelles permet supprimer les points d'intérêt qui ne sont pas stables (Figure 130 a) et c)).

Comme nous allons en discuter dans la section 5.3, seulement quatre points sont nécessaires pour estimer la pose. Tous les détecteurs présentés sont donc potentiellement utilisables dans notre application. La condition est d'être capable de mettre en correspondance correctement les points répétés. Généralement, pour cela, un descripteur de points d'intérêt est utilisé. Cette étape est expliquée dans la section suivante (section 5.2.2). La mise en correspondances des points d'intérêt comporte des erreurs (également appelées *outliers* dans la littérature anglophone). En pratique, entre une image de synthèse éloignée de moins de 10° de point de vue avec une image réelle de face, la méthode *DoG* détecte 60 points d'intérêt correctement mis en correspondance sur 70 points au total. Cette situation est favorable car la vue de face du visage possède suffisamment de texture. Par contre pour une vue orientée à ~30° en *yaw*, seulement 20 points d'intérêt sont correctement mis en correspondance sur 25 points au total. Il est important de veiller à conserver un nombre de points d'intérêt correctement mis en correspondance quelle que soit l'orientation du visage. Maximiser la répétabilité du détecteur (en taux et en nombre total) fait partie de cette démarche.

Le détecteur *DoG* est adapté à notre application. Cependant, il est souvent considéré comme lent et incompatible avec des applications en temps réel. Un détecteur qui n'est pas multi-échelles, comme *Harris*, est plus rapide et peut être utilisé. Cependant le faible taux de répétabilité de ce type de détecteur engendre inévitablement des erreurs de mis en correspondance. Pour une vue orientée à ~30° en *yaw* 17 points d'intérêt sont correctement mis en correspondance sur 35 au total avec le détecteur de *Harris*. Les aspects temps réels n'étant pas l'axe majeur de ces recherches, dans la suite nous choisissons de travailler avec le descripteur *DoG* car il est le plus performant.

5.2.2. Les descripteurs de points d'intérêt

Un descripteur permet de discriminer les points d'intérêt. L'objectif est de pouvoir les reconnaître quand ils apparaissent sur deux images d'une même scène prise dans des conditions différentes (bruit, point de vue, illumination). Un détecteur performant doit être robuste face à ces variations. Le descripteur le plus simple est un vecteur composé des valeurs des pixels qui constituent l'entourage du point d'intérêt. Cependant, cette méthode n'est pas robuste aux variations. Les descripteurs les plus performants sont inspirés de la vision humaine.

Nous allons présenter et utiliser dans ce manuscrit uniquement la méthode implémentée dans l'algorithme *SIFT* de Lowe [87]. Il s'agit de la description, utilisant les histogrammes des orientations, représentée sur la Figure 131. Lowe propose de décrire l'amplitude et l'orientation du gradient. L'orientation et l'amplitude du gradient sont calculées dans une fenêtre de 16 × 16 pixels autours du point détecté. De cette zone, 4 × 4 cellules sont formées à partir de 4 × 4 pixels. Dans chaque sous-groupe va être réalisé un histogramme des orientations. Celui-ci est composé de huit classes, chacune d'elles représentant une orientation dans l'intervalle [0° 360°]. Chaque pixel d'une cellule contribue à l'une des huit classes d'orientation avec un poids relatif à l'amplitude du gradient. La zone de 16 × 16 pixels est préalablement multipliée par une gaussienne d'échelle $\sigma = 8\,pixels$ afin que les pixels les plus éloignés du point d'intérêt

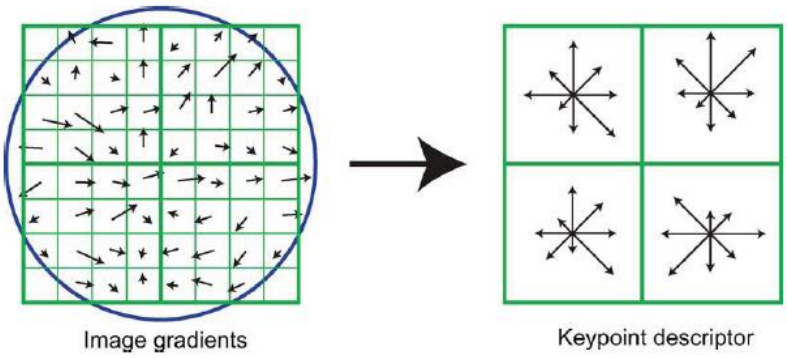


Figure 131. Descripteur implémenté dans l'algorithme SIFT [87]. Dans cet exemple, un voisinage de 8 × 8 pixels est considéré (au lieu de 16 × 16 dans l'algorithme original) autour du point d'intérêt (à gauche du schéma). 2 × 2 histogrammes (au lieu de 4 × 4 dans l'algorithme original) sont évalués. Il s'agit des cellules (à droite du schéma). Une cellule est composée de 4 × 4 pixels.

aient un poids plus faible (cette gaussienne est symbolisée par le cercle bleu sur la Figure 131). La discrétisation des orientations rend le descripteur *SIFT* robuste aux faibles changement de points de vue.

Remarque : la technique portant le nom d'histogramme de gradients orientés (HoG pour histogram of oriented gradients) proposé Dalal et Triggs [156] est relativement proche de celle proposée dans la méthode SIFT par Lowe. Dans un contexte de reconnaissance de piétons, ces derniers proposent de calculer le HoG sur toute l'image selon une grille dense. Puis, grâce à un classifieur supervisé, il a été montré que la détection et la localisation du piéton sur l'image peuvent être réalisées avec un bon taux de performance.

5.3. Déduction de la pose à partir des correspondances 3D-2D

Dans la section 3.5.3 une base d'images de synthèse 2D du visage du conducteur a été créée. Dans la section 5.2.1.5, il a été montré comment il était possible d'approximer les coordonnées 3D des points d'intérêt détectés sur les images de synthèse. Les descripteurs permettent ensuite de mettre en correspondance des points d'intérêt détectés sur une image questionnée avec les points d'intérêt des images de synthèse. Grâce à ces correspondances 3D-2D, il est possible d'estimer la pose relative entre la caméra et le modèle 3D. L'objectif de cette section est de formuler mathématiquement le problème de recherche de la pose à partir des correspondances 3D-2D. Quelques-unes des méthodes les plus populaires de littérature seront présentées et testées dans le cas spécifique de l'estimation de la pose du visage par caméra thermique.

5.3.1. Le problème *PnP* (Perspective-n-Point)

L'équation 3.11 modélise le processus complet de formation de l'image à partir des coordonnées 3D d'un objet dans le repère monde et des paramètres extrinsèques ($[R|t]$) et intrinsèques (K) de la caméra.

Le problème *perspecive-n-point (PnP)*, dont la formulation est connue depuis 1841 [157], a été abordé par la communauté de vision par ordinateur par Fischler et Bolles en 1981 [158]. L'objectif est de rechercher la transformation, représentée par la matrice paramètres extrinsèques, qui permet de passer du repère monde au repère caméra. Il est supposé qu'un ensemble de *n* correspondances 3D-2D ait été préalablement établi.

En d'autres termes, une étape préalable permet d'associer des points détectés sur un objet 3D exprimés dans le repère monde, avec leurs projections 2D exprimées dans le plan image de la caméra. La plupart des méthodes *PnP* consistent à exprimer les points 3D dans le référentiel de la caméra. Il est ensuite possible de trouver la pose qui permet d'aligner le repère monde et le repère caméra [159]. Dans la formulation du problème *PnP*, il est assumé que les paramètres intrinsèques de la caméra sont connus.

5.3.1.1. Le problème *P3P*

Le plus petit ensemble de points nécessaire à la résolution du problème *PnP* est $n = 3$. Dans ce cas, le problème d'estimation est appelé *perspective-3-point* (*P3P*). Le problème *P3P* se résume à trouver les racines d'un polynôme de degré 4 [160]. Un quatrième point, au minimum, est nécessaire pour lever l'incertitude et trouver la solution parmi les quatre racines du polynôme.

Formulons le *P3P* à partir de la géométrie illustrée sur la Figure 132. Soient trois points $\boldsymbol{Q}_i$ tel que $i = 1,2,3$. Ces points sont souvent appelés *control points* ou *reference points* dans la littérature anglophone. Nous connaissons leurs coordonnées 3D exprimées dans le repère objet car nous disposons d'un maillage 3D de l'objet observé. Elles sont notées $\boldsymbol{Q}_i^O$. Les inter-distances (a, b et c) entre les trois points de contrôle sont ainsi connues. Nous connaissons également les coordonnées (x_i, y_i) des projections $\boldsymbol{q}_i$ tel que $i = 1,2,3$ dans le plan image. Nous cherchons les distances entre les $\boldsymbol{Q}_i$ et le centre de projection de la caméra, c'est-à-dire les distance $d_i = \left\|\overrightarrow{\boldsymbol{O}_C\boldsymbol{Q}_i}\right\|$, telles que $i = 1,2,3$. Les distances d_i sont souvent appelées profondeurs. Lorsque les profondeurs des points de contrôle sont déterminées, les cordonnées 3D sont

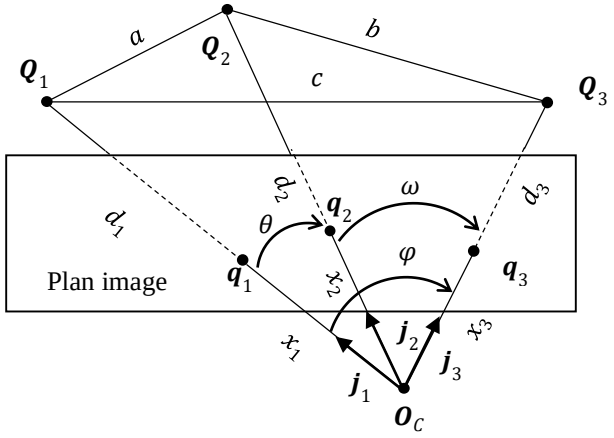


Figure 132. Illustration de la géométrie de trois points 3D ($\boldsymbol{Q}_1, \boldsymbol{Q}_2$ *et* $\boldsymbol{Q}_3$) et du centre de projection (O_C). Les longueurs des côtés du triangle formé par les points 3D sont connues (a, b *et* c). Le problème est de trouver les distances ($d_1 = \left\|\overrightarrow{\boldsymbol{O}_C\boldsymbol{Q}_1}\right\|, d_2 = \left\|\overrightarrow{\boldsymbol{O}_C\boldsymbol{Q}_2}\right\|$ *et* $d_3 = \left\|\overrightarrow{\boldsymbol{O}_C\boldsymbol{Q}_3}\right\|$). Elles permettent de déterminer complétement les points 3D. La distance focale est supposée connue ainsi que les positions des projections dans le plan image ($\boldsymbol{q}_1, \boldsymbol{q}_2$ *et* $\boldsymbol{q}_3$).

entièrement connues dans le repère caméra (on les notes $\boldsymbol{Q}_i^C$). Le problème revient alors à chercher la matrice de rotation $\boldsymbol{R}$ et le vecteur de translation $\boldsymbol{t}$ qui lient les coordonnées des points de contrôle 3D exprimés dans le repère objet à leurs coordonnées 3D exprimées dans le repère caméra.

$$\boldsymbol{Q}_i^C = \boldsymbol{R}\boldsymbol{Q}_i^O + \boldsymbol{t} \tag{5.15}$$

Plusieurs méthodes permettent de trouver les paramètres extrinsèques $[\boldsymbol{R}|\boldsymbol{t}]$ à partir des coordonnées 3D dans le repère caméra et objet [161, 162].

Pour trouver les d_i, il faut utiliser les angles θ, φ et ω définis comme suit :

$$\theta = \widehat{Q_1 O_C Q_2},$$

$$\varphi = \widehat{Q_1 O_C Q_3},$$

$$\omega = \widehat{Q_2 O_C Q_3}.$$

Ensuite, il faut exprimer les vecteurs unitaires qui pointent vers les Q_i, notés $\boldsymbol{j}_i$ tel que $i = 1,2,3$. Ils sont définis comme suit :

$$\boldsymbol{j}_i = \frac{1}{\sqrt{x_i + y_i + f}} \begin{bmatrix} x_i \\ y_i \\ f \end{bmatrix}, \qquad i = 1,2,3 \tag{5.16}$$

La distance focale est notée f. Il est ainsi possible d'exprimer les angles en question, en fonction des $\boldsymbol{j}_i$:

$$\begin{aligned} cos\,\theta &= \boldsymbol{j}_1.\boldsymbol{j}_2, \\ cos\,\varphi &= \boldsymbol{j}_1.\boldsymbol{j}_3, \\ cos\,\omega &= \boldsymbol{j}_2.\boldsymbol{j}_3. \end{aligned} \tag{5.17}$$

Finalement, la loi des cosinus permet d'écrire un système de trois équations à trois inconnues :

$$\begin{cases} d_1^2 + d_2^2 - 2d_1 d_2\, cos\,\theta = a^2 \\ d_1^2 + d_3^2 - 2d_1 d_3\, cos\,\varphi = b^2 \\ d_2^2 + d_3^2 - 2d_2 d_3\, cos\,\omega = c^2 \end{cases} \tag{5.18}$$

Un grand nombre de méthodes propose de réduire ce système en une seule équation d'ordre 4.

$$A_4 v^4 + A_3 v^3 + A_2 v^2 + A_1 v + A_0 = 0 \tag{5.19}$$

Une fois la variable v trouvée, il est possible de trouver les profondeurs des points de contrôle (les d_i). Haralick & *al.* listent et comparent six différentes manières de passer de l'équation (5.18) à l'équation (5.19) [163]. Ces six méthodes ont deux défauts en commun. Le premier est l'instabilité numérique du résultat en fonction de l'ordre dans lequel les trois points de contrôle sont traités. Le second est l'incapacité de trouver une solution dans certaines configurations géométrique comme la configuration '*danger cylinder'*.

Li & Xu proposent de reformuler le problème *P3P* en un problème qu'ils appellent *perspective similar triangle (PST)* [164]. Leur formulation apparait plus robuste à la fois au problème de permutations des points de contrôle et aux singularités géométriques. L'idée principale de leur formulation consiste à rechercher les profondeurs (d_i') de trois nouveaux points de contrôle notés $\boldsymbol{Q}_i'$ positionnées sur un triangle semblable à celui formé par les trois points de contrôle initiaux que sont les $\boldsymbol{Q}_i$. En posant que $d_i' = 1$ le problème revient à rechercher d_2' et d_3' et un facteur λ qui traduit la proportionnalité $d_i = \lambda d_i'$. Finalement,

le problème revient encore à rechercher les racines d'un polynôme d'ordre 4, mais la stabilité numérique est meilleure dans ce cas.

Une autre spécificité importante de leur méthode concerne la manière de rechercher les racines du polynôme. Selon Li & Xu, le bruit de mise en correspondance affecte le calcul des coefficients du polynôme. Ainsi, il est possible de manquer des racines (cf. Figure 133). Ils proposent d'explorer les extrema du polynôme d'ordre 4. Soit F le polynôme résultant du problème *PST* ou *P3P*.

$$F(v) = A_4 v^4 + A_3 v^3 + A_2 v^2 + A_1 v + A_0 \tag{5.20}$$

Sa dérivée F' est un polynôme d'ordre 3. En l'étudiant, il est possible d'extraire au maximum trois extrema notés v_i avec $i = 1,2,3$. Si v_i' est un maximum et $F(v_i) < 0$ alors v_i est potentiellement une racine manquée. Il en est de même si v_i' est un minimum et $F(v_i) > 0$.

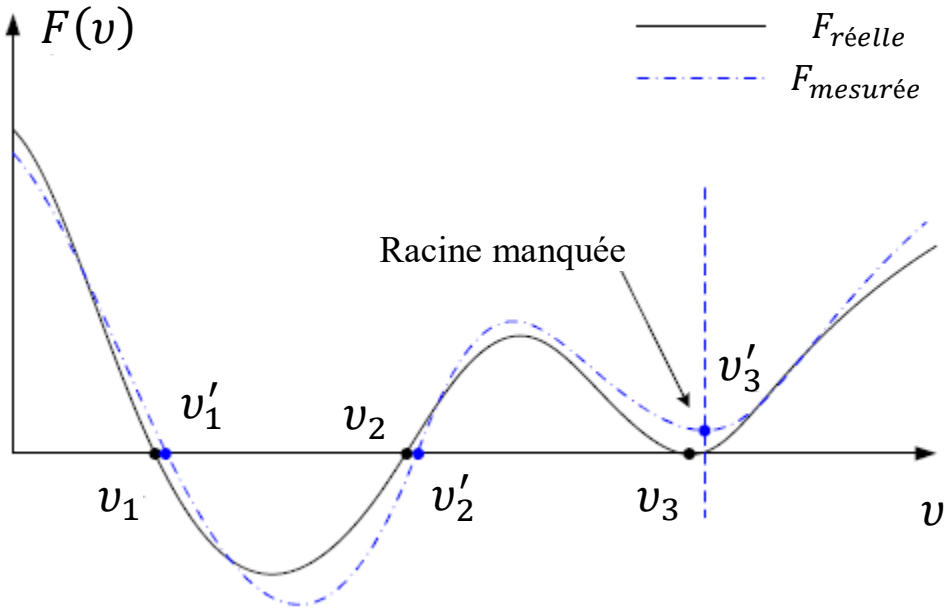


Figure 133. Illustration de l'impact du bruit sur la résolution du problème P3P..

5.3.1.2. Le problème *PnP* à partir de plusieurs *P3P*

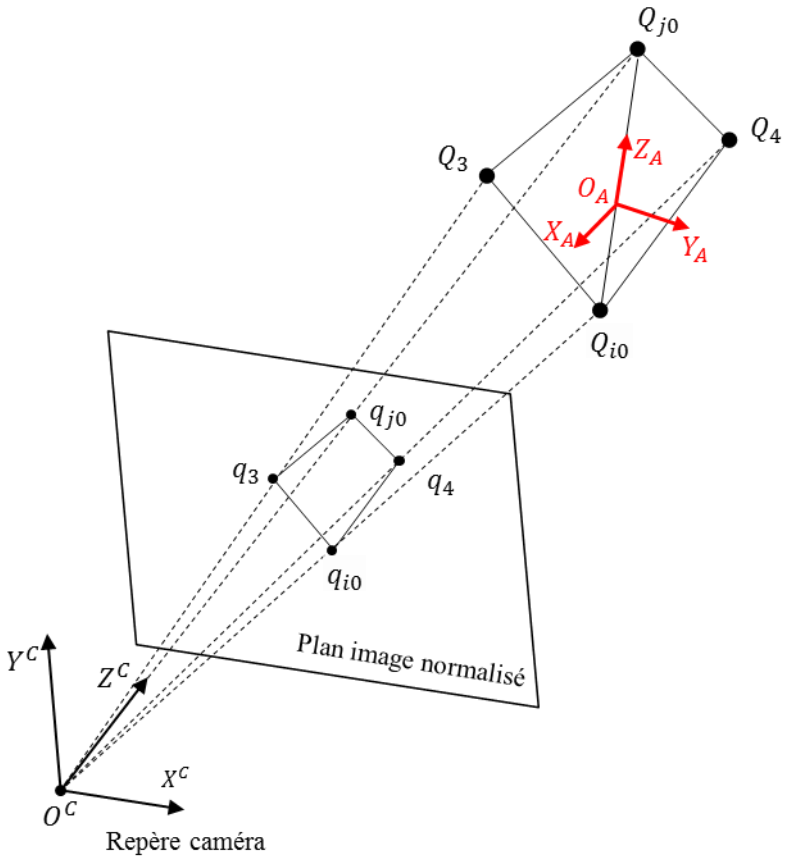


Figure 134. Changement de repère dans la méthode de Li & Xu.

En pratique le nombre de correspondances *n* est supérieur à trois. Il s'agit alors d'un problème *PnP*. Cette section explique une méthode permettant d'en tirer profit. Généralement, le problème *PnP* est divisé en plusieurs problèmes *P3P*. Ainsi, pour *n* correspondances, il est possible d'établir $\frac{(n-1)(n-2)}{2}$ polynômes du quatrième degré. Certaines méthodes permettent de résoudre un tel système non-linéaire par des techniques de linéarisation [160].

Li & Xu ont également proposé une approche différente [165]. Leur méthode est intitulée *RPnP* pour *robust PnP*. Elle est couplée à leur formulation *PST* du problème *P3P*. Les *n* $\boldsymbol{Q}_i$ vont être exprimés, non plus dans le repère objet, mais dans un nouveau repère noté $\boldsymbol{O^A X^A Y^A Z^A}$. Ce nouveau repère est illustré sur Figure 134. Il est défini comme suit. En premier lieu, il faut déterminer le vecteur $\boldsymbol{Z}_A$. Pour ce faire, *n* inter-distances entre les $\boldsymbol{Q}_i$, choisies au hasard parmi les $\binom{n}{2}$ possibles, sont considérées. L'inter-distance la plus grande, notée $\overline{\boldsymbol{Q}_{i0}\boldsymbol{Q}_{j0}}$, est retenue et l'axe Z_a est portée par $\overrightarrow{\boldsymbol{Q}_{i0}\boldsymbol{Q}_{j0}}$. L'origine $\boldsymbol{O}_A$ de ce nouveau repère est placée au centre de $\overrightarrow{\boldsymbol{Q}_{i0}\boldsymbol{Q}_{j0}}$. Le problème d'estimation de la pose revient désormais à rechercher la matrice de rotation et le vecteur de translation liant le repère caméra au repère $\boldsymbol{O^A X^A Y^A Z^A}$.

Ensuite *n – 2* sous-ensembles de trois points $\{\boldsymbol{Q}_{i0}\boldsymbol{Q}_{j0}\boldsymbol{Q}_k\}$ avec $k = \{1,2,\dots,n-2\}$ vont être formés. *n – 2* polynômes sont déduits de ces trios impliquant toujours les points $\boldsymbol{Q}_{i0}$ et $\boldsymbol{Q}_{j0}$. On obtient le système d'équations suivant :

$$\begin{cases} F_1(v) = a_1v^4 + b_1v^3 + c_1v^2 + d_1v + e_1 = 0 \\ F_2(v) = a_2v^4 + b_2v^3 + c_2v^2 + d_2v + e_2 = 0 \\ \dots \\ F_{n-2}(v) = a_{n-2}v^4 + b_{n-2}v^3 + c_{n-2}v^2 + d_{n-2}v + e_{n-2} = 0 \end{cases} \tag{5.21}$$

La résolution de ce système permet de connaître les coordonnées 3D des points $\boldsymbol{Q}_{i0}$ et $\boldsymbol{Q}_{j0}$ dans le repère caméra. Pour le résoudre, le parti pris de Li & Xu est d'explorer les minima de la somme quadratique (celle-ci est notée $\mathcal{F}$) des *n - 2* polynômes afin de ne pas manquer de racines comme cela a été abordé au paragraphe précédent pour la recherche des racines d'un seul polynôme.

$$\mathcal{F} = \sum_{i=1}^{n-2} F_i^2(v) \tag{5.22}$$

Une fois le système (5.21) résolu, $\boldsymbol{Z}_A$ peut être exprimé dans le repère caméra. Ainsi, la matrice de rotation qui permet de passer de $\boldsymbol{O}^A\boldsymbol{X}^A\boldsymbol{Y}^A\boldsymbol{Z}^A$ à $\boldsymbol{O}^C\boldsymbol{X}^C\boldsymbol{Y}^C\boldsymbol{Z}^C$ peut être décomposée comme suit :

$$\boldsymbol{R} = \boldsymbol{R}'rot(Z,\alpha) = \begin{bmatrix} r_1 & r_4 & r_7 \\ r_2 & r_5 & r_8 \\ r_3 & r_6 & r_9 \end{bmatrix}\begin{bmatrix} c & -s & 0 \\ s & c & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{5.23}$$

avec $c = \cos(\alpha)$, $s = \sin(\alpha)$ et,

$$\begin{aligned} \boldsymbol{R}' = \begin{bmatrix} r_1 & r_4 & r_7 \\ r_2 & r_5 & r_8 \\ r_3 & r_6 & r_9 \end{bmatrix} &= rot(X,\theta)rot(Y,\varphi) \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\theta) & -sin(\theta) \\ 0 & sin(\theta) & cos(\theta) \end{bmatrix}\begin{bmatrix} cos(\varphi) & 0 & sin(\varphi) \\ 0 & 1 & 0 \\ -sin(\varphi) & 0 & cos(\varphi) \end{bmatrix}, \end{aligned} \tag{5.24}$$

et

$$\begin{aligned} \theta &= acos\left(\frac{\boldsymbol{Z}_A.\boldsymbol{X}_C}{\|\boldsymbol{Z}_A\|\|\boldsymbol{X}_C\|}\right), \\ \varphi &= acos\left(\frac{\boldsymbol{Z}_A.\boldsymbol{Y}_C}{\|\boldsymbol{Z}_A\|\|\boldsymbol{Y}_C\|}\right). \end{aligned} \tag{5.25}$$

La projection des points 3D des points du repère $\boldsymbol{O}^A\boldsymbol{X}^A\boldsymbol{Y}^A\boldsymbol{Z}^A$ sur le plan image normalisé (normalisation par rapport aux paramètres intrinsèques) s'exprime comme suit :

$$\lambda\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} r_1 & r_4 & r_7 \\ r_2 & r_5 & r_8 \\ r_3 & r_6 & r_9 \end{bmatrix}\begin{bmatrix} c & -s & 0 \\ s & c & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{5.26}$$

Finalement, en réarrangeant les termes, Li & Xu se ramènent à un problème de décomposition en valeurs singulières avec $n \geq 4$ car nous traitons le problème *PnP*.

$$[A_{2n\times 1} \quad B_{2n\times 1} \quad C_{2n\times 4}]\begin{bmatrix} c \\ s \\ t_x \\ t_y \\ t_z \\ 1 \end{bmatrix} = 0 \tag{5.27}$$

La robustesse de cette méthode tient avant tout, d'après ses auteurs, de la recherche des minima de $\mathcal{F}$ et de la manière de rechercher deux rotations en donnant plus d'importance aux couples de points les plus éloignés. De plus il est nécessaire d'évoquer le fait que, contrairement à d'autres méthodes, elles aussi non-itératives, la complexité de calcul augmente de façon linéaire avec le nombre de points (complexité de calcul en $O(n)$).

Remarque n°1 : les travaux de Li & Xu sont assez proches des travaux de Lepetit [166].

Remarque n°2 : nous avons discuté des méthodes PnP non-itératives. Il existe une partie de la littérature qui traite des méthodes itératives. Ces dernières sont considérées précises mais nécessitent beaucoup plus de calculs. De plus il est nécessaire de les initialiser correctement.

5.3.2. Prise en compte des erreurs de mise en correspondance

Lorsque deux points sont mis en correspondance alors qu'ils n'appartiennent pas à la même zone de l'objet alors, il s'agit d'une erreur. Le terme *outliers* est utilisé dans la littérature anglophone. Par opposition, les correspondance correctes sont appelées *inliers*. Lorsque deux points sont mis en correspondances mais la localisation n'est pas précise on parle alors de bruit de mise en correspondance. Il n'existe pas une limite clairement établie permettant de distinguer le bruit des *outliers*. Cependant, dans la littérature le bruit est généralement étudié jusqu'à une erreur de 20 pixels (écart type d'une gaussienne).

Même si une attention particulière est portée sur le choix du détecteur et du descripteur pour les minimiser, ces erreurs sont inévitables. Une méthode bien connue dans le domaine de la vision par ordinateur permet d'y être robuste. Il s'agit de *RANSAC* (*random sample consensus* dans la littérature anglophone). Applicable à d'autres types de problème comportant des *outliers*, cette méthode fut justement développée par Fisher & Bolles [158] dans un contexte *PnP*. *RANSAC* est une méthode qui va itérativement prélever au hasard un jeu de données pour tenter de décrire au mieux l'ensemble des données. Dans le cas *PnP*, lors d'une itération, *RANSAC* sélectionne au moins quatre correspondances 3D-2D (le nombre de correspondances est un paramètre d'entrée). On en déduit une estimation de la pose. Puis, l'erreur de reprojection utilisant cette estimation est calculée. On établit ensuite le pourcentage d'*inliers*, c'est-à-dire le nombre de correspondances dont l'erreur de reprojection est inférieure à un seuil (typiquement 10 pixels) sur le nombre total de correspondances. L'itération suivante est considérée meilleure, si ce pourcentage est dépassé. Dans ce cas le pourcentage d'*inliers* à dépasser est mis à jour et on passe à l'itération suivante. Lorsque le pourcentage d'*inliers* atteint un seuil, fixé par l'urilisateur, l'algorithme *RANSAC* a terminé. Nous discuterons du choix du nombre d'itérations *RANSAC* à la fin de la section 5.4.2.

Concernant le temps de calcul, en 2014 Ferraz & *al.* ont proposé de traiter le rejet des *outliers* dans le contexte *PnP* [167] sans utiliser la méthode *RANSAC*. Leur méthode est nommée *REPPnP* (*robust efficient*

procrustes PnP). Elle pourrait être avantageuse car elle permet de conserver un temps de calcul approximativement constant en fonction du pourcentage d'*outliers* et du nombre de points total.

Il s'avère que cette méthode n'est pas stable dans notre cas. Pour le constater, nous avons utilisé deux images de la base de modèles d'angle *pitch* identique et espacées de 5° selon l'angle *yaw*. Bien que ces deux images soient très similaires, un taux d'*outliers* non négligeable est déjà présent. Il y a également un bruit de mise en correspondance. Cette combinaison caractéristique de l'imagerie thermique ne permet pas à l'algorithme *REPPnP* d'estimer la pose.

Dans la méthode *RPnP* de Xu & *al.*, le temps de calcul progresse linéairement avec le nombre de points. En moyenne nous utilisons quarante points. Or Ferraz & *al.* font remarquer que la méthode *RPnP* est la méthode la plus rapide lorsque le nombre de correspondances est inférieure à 100 et qu'il n'y a pas d'*outliers*.

La méthode *RPnP* est la mieux adaptée à notre besoin car c'est la seule capable d'estimer la pose dans un contexte de mise en correspondance difficile. La réduction du taux d'*outliers* semble un paramètre clefs permettant d'accélérer le temps de calcul. Nous en discuterons lors des détails de l'implémentation exposés dans la section suivante.

5.4. Détails de l'implémentation

5.4.1. Principe général

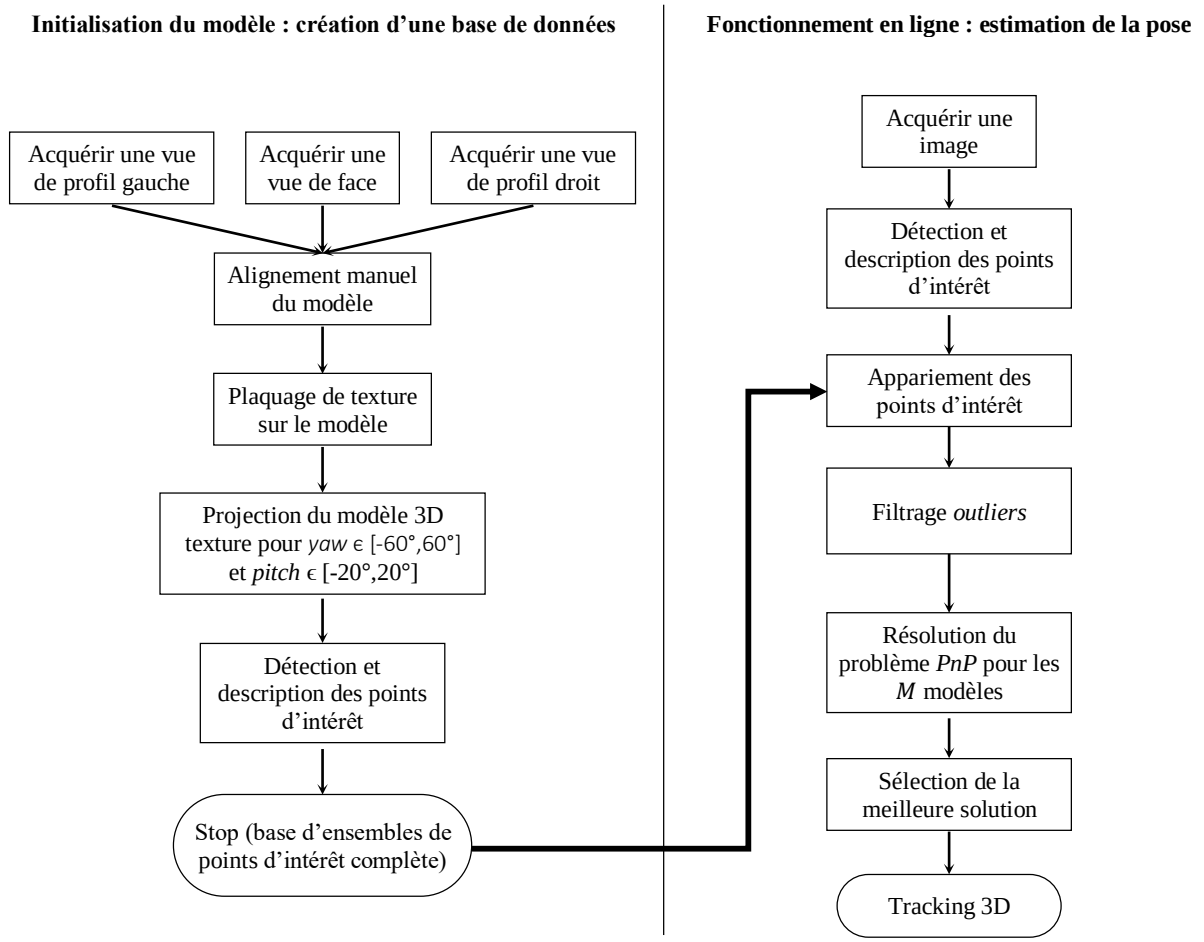


Figure 135. Principe générale d'estimation de la pose. Sur la partie de gauche on retrouve les étapes 'hors ligne' de plaquage de texture et de création de la base d'images de synthèse détaillées au chapitre 3. Sur la partie de droite, nous expliquons le principe d'estimation de la pose 'en ligne'.

Le principe général de notre méthode est illustré sur la Figure 135. Sur la partie gauche de cette figure, on retrouve l'étape d'initialisation du modèle 3D décrite au chapitre 3.

Sur la partie droite de cette figure, le fonctionnement 'en ligne' est décrit. Les étapes de **détection et de description des points d'intérêt** ont été abordées précédemment.

Les détails pratiques d'**appariement de points d'intérêt** (c'est-à-dire la mise en correspondances 3D-2D) vont être exposés. De plus une étape de **filtrage des *outliers*** (avant l'étage *PnP* qui intègre un rejet des *outliers* grâce à la méthode *RANSAC*) va être détaillée.

La **résolution du problème *PnP*** à partir de correspondances 3D-2D a déjà été décrite précédemment. Sur la Figure 135, il est précisé que le problème *PnP* est résolu pour M modèles de la base d'images de

synthèse, c'est-à-dire pour M ensembles de correspondances 3D-2D. En pratique nous utilisons M=225 images de synthèse. Nous montrerons l'influence d'une réduction de M sur la précision de l'estimation de l'orientation dans la section 5.4.3. Dans cette même section, il sera expliqué que le critère permettant de choisir la meilleure solution parmi les M est basé sur la minimisation de l'erreur de reprojection ainsi que sur la maximisation du nombre de correspondances.

5.4.2. Mise en correspondance 3D-2D

5.4.2.1. Principe général

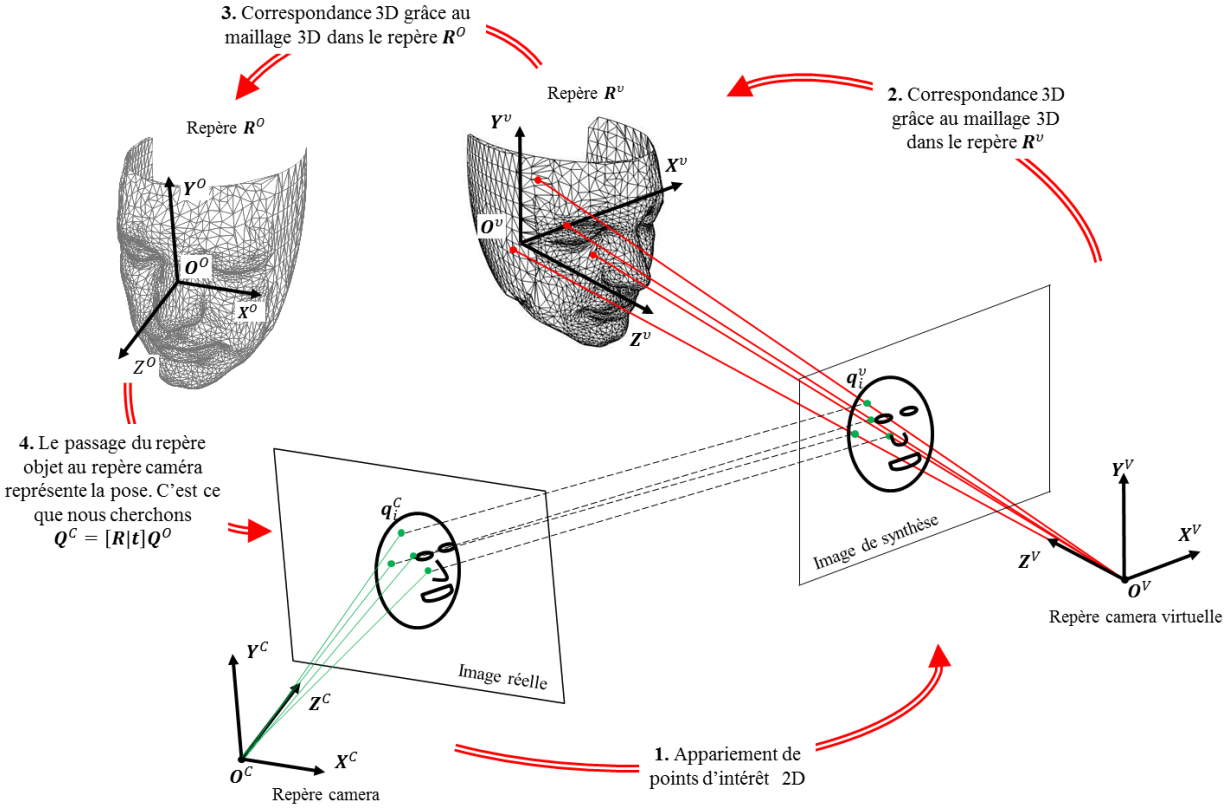


Figure 136. Principe de mise en correspondance 3D-2D. (1) Mise en corresponance de points d'intérêt entre l'image réelle et une image de synthèse. (2) correspondance 3D-2D entre l'image de synthèse et le modèle 3D. (3) Recherche de la pose entre le repère caméra et le repère objet fixe.

Comme déjà annoncé, nous utilisons une base d'images de synthèse (la base de modèles) pour obtenir les correspondances en coordonnées 3D des points d'intérêt détectés en 2D sur l'image réelle acquise par la caméra. Les trois premières étapes illustrées sur la Figure 136 et détaillées ci-dessous permettent d'expliquer le processus de mise en correspondance 3D-2D.

1. Considérons une image réelle et une image de synthèse. Reprécisons que l'image de synthèse a été obtenue en projetant les *vertices* $\boldsymbol{Q}_k$ du maillage 3D partir de l'équation suivante :

$$\boldsymbol{q}_k^v = \boldsymbol{K}.\boldsymbol{R}^v.\boldsymbol{Q}_k^O$$

Grâce à la méthode *SIFT* il est possible de mettre en correspondance un certain nombre de points d'intérêt. Les points d'intérêt détectés sur l'image réelle sont notés $\boldsymbol{q}_i^c$. Ceux détectés sur l'image de synthèse sont notés $\boldsymbol{q}_i^v$.

2. On recherche la projection du maillage 3D $\boldsymbol{q}_k^v$ la plus proche du point $\boldsymbol{q}_i^v$ détecté sur l'image de synthèse (cf. équation (5.13)). Cette recherche constitue une approximation. Les coordonnées 3D $\boldsymbol{Q}_k^v$ dans le repère $\boldsymbol{O^v X^v Y^v Z^v}$ du *vertexe* associées à $\boldsymbol{q}_k^v$ sont connues. On considère qu'elles approximent les coordonnées 3D des points $\boldsymbol{q}_i^v$ détectés sur l'image de synthèse, on les note donc $\widehat{\boldsymbol{Q}_i^v}$.

3. Pour décrire les points 3D, il est préférable d'utiliser le repère $\boldsymbol{O^O X^O Y^O Z^O}$ plutôt que le repère $\boldsymbol{O^v X^v Y^v Z^v}$ afin d'éviter tout problème lié à une erreur de positionnement des axes de rotation (cf. section 3.5.1.3). Le passage entre ces deux repères étant tout à fait connu, il est aisé de connaitre les coordonnées 3D dans le repère $\boldsymbol{O^O X^O Y^O Z^O}$. Ces coordonnées sont notées $\widehat{\boldsymbol{Q}_i^O}$

4. Finalement les correspondances $\widehat{\boldsymbol{Q}_i^O} \leftrightarrow \boldsymbol{q}_i^c$ alimentent l'algorithme de résolution du problème *PnP* de Li & Xu.

Remarque : Les coordonnées 3D sont exprimées dans un repère fixe (le repère $\boldsymbol{O^O X^O Y^O Z^O}$). Il est donc implicitement considéré que la caméra est en mouvement relatif par rapport au repère fixe $\boldsymbol{O^O X^O Y^O Z^O}$. Une fois de plus nous faisons remarquer qu'il est équivalent de considérer la caméra en mouvement relatif et l'objet 3D fixe que de considérer la caméra fixe et l'objet en mouvement relatif.

5.4.2.2. Réduction du nombre d'*outliers*

Utilisation du niveau continu *NC* du voisinage local :

Concernant l'étape **1**. du plan ci-dessus, apportons quelques précisions. Bien que l'algorithme *SIFT* soit performant, en pratique, il y a un certain nombre d'erreurs de mise en correspondance. L'un des atouts de la description dans l'algorithme *SIFT* (et de bien d'autres méthodes d'appariement) est sa capacité à mettre en correspondance des points d'intérêt sans prendre en compte le niveau continu *NC* de la zone de pixels concernée (cf. 5.2.2). En imagerie visible cela permet d'être robuste aux variations d'illumination. En imagerie thermique cette robustesse peut sembler moins utile puisque l'on détecte seulement le rayonnement propre des objets. Ainsi dans la littérature, certains proposent d'utiliser le niveau continu *NC* moyen de la zone d'intérêt pour supprimer les *outliers* [101]. L'information du *NC* du voisinage local, appelé *AI* (*averaging information*), s'exprime comme suit :

$$AI(Y, i_0, j_0) = \frac{1}{N}\sum_{ij} Y(i,j) \tag{5.28}$$

L'image est notée Y et les coordonnées du point d'intérêt sont notées i_0, j_0. La fenêtre du voisinage local considéré comporte N valeurs. Si la différence des niveaux moyens des voisinages locaux des points mis en correspondance dépasse un certain seuil, la mise en correspondance est rejetée et considérée comme un *outlier*. Dans notre cas et avec nos notations, cette condition s'exprime comme suit :

$$AI(Y, i_1, j_1) - AI(g_c, i_2, j_2) = \begin{cases} < seuil, \ \ acceptée\ (inlier) \\ > seuil, \ \ rejetée\ (oulier) \end{cases} \quad (5.29)$$

On détecte un point d'intérêt aux coordonnées i_1, j_1 de l'image réelle notée Y. Celui-ci est mis en correspondance grâce à l'algorithme *SIFT* au point d'intérêt situé aux coordonnées i_2, j_2 de la $c^{ième}$ image de synthèse notée g_c.

Les auteurs utilisent cette méthode pour la reconnaissance de visages. Plusieurs formes de fenêtre ont été explorées et donnent lieu à plusieurs noms anglophones : *SWF* pour *star-styled window filter* et *YWF* pour *Y-styled window filter*. Comme l'invariance par rotation dans le plan n'est pas nécessaire dans leur application, une fenêtre en forme de *Y* permet de filtrer plus d'erreur de mise en correspondance qu'une fenêtre circulaire ou en forme d'étoile. Dans notre cas, nous souhaitons être robuste aux rotations dans le plan donc nous avons opté pour une fenêtre en étoile (*SWF*).

Il est toutefois nécessaire de prendre des précautions concernant l'ajout du niveau continu *NC* dans le descripteur *SIFT*. Comme cela a été présenté dans le chapitre 2, le *NC* varie en fonction de la température de la caméra. Seul un étalonnage radiométrique grâce à un élément extérieur étalonné ou grâce à un étalonnage en chambre climatique (cf. section 2.6) permet d'envisager ce type de filtrage.

Utilisation de la norme et de la direction du « vecteur d'appariement » :

Si l'on considère qu'il est possible d'ajouter le *NC* dans la description des points d'intérêt, un pourcentage résiduel très faible d'*outliers* demeure. Il est possible d'en supprimer une partie supplémentaire. Pour cela nous considérons que les « vecteurs d'appariement », qui relient les points d'intérêt entre l'image réelle et l'image de synthèse, doivent approximativement avoir la même norme et la même direction. Ce filtrage fonctionne correctement seulement dans des cas de faibles rotations dans le

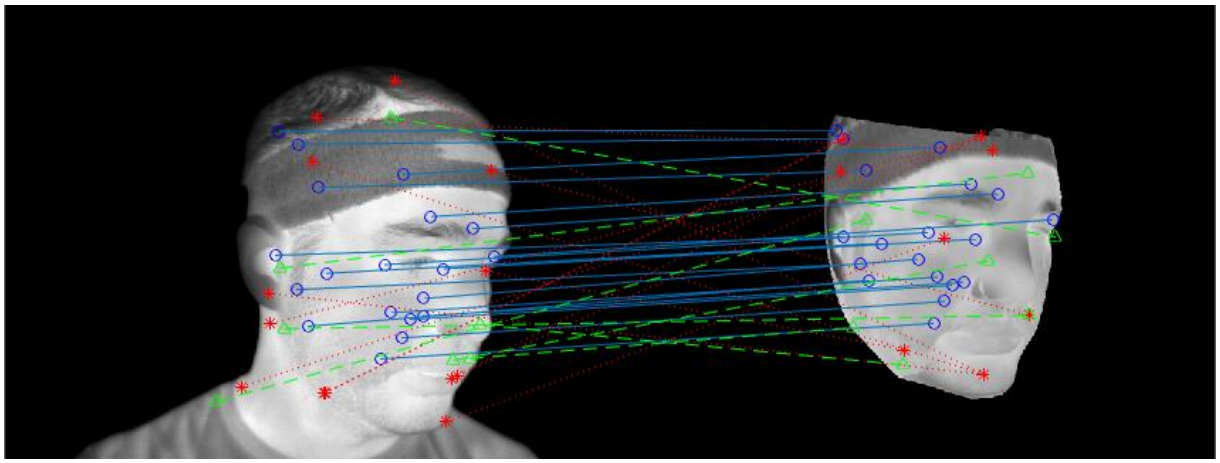


Figure 137. Illustration de l'étape de rejet des outliers. Les inliers sont représentés par o—o. Les outliers filtrés par la méthode *SWF* sont représentés par *..*. Les outliers filtrés par la norme et la direction du vecteur d'appariement sont reprsentés par Δ--Δ.

plan, ce qui est le cas dans notre application. La Figure 137 illustre l'effet des filtrages *SWF* et sur le norme et la direction du vecteur d'appariement.

La réduction du pourcentage d'*outliers* est importante car elle permet à l'algorithme *RANSAC* de converger rapidement. On peut ainsi limiter le nombre d'itération. Pour s'en convaincre réalisons l'expérience suivante. Considérons une image réelle du visage et une image de synthèse. Dans l'image réelle, les orientations de la pose du visage valent *yaw*=38° et *pitch*=11.5° d'après la centrale inertielle. Nous choisissons une image de synthèse labellisée avec un angle *yaw*=35° et un angle *pitch*=10°. Nous allons détecter et mettre en correspondance des points d'intérêt. Dans un premier cas, nous utiliserons le filtrage *SWF* avec un seuil (cf. équation (5.29)) valant $\boldsymbol{seuil_{SWF} = 320\ Adu}$ **(sur 16 bits)** et le filtrage sur la norme et la direction des vecteurs de mise en correspondance. Dans un second cas, nous n'utilisons aucun filtrage. Nous estimons la pose grâce à l'algorithme *RPnP* combiné à la méthode *RANSAC* pour le filtrage

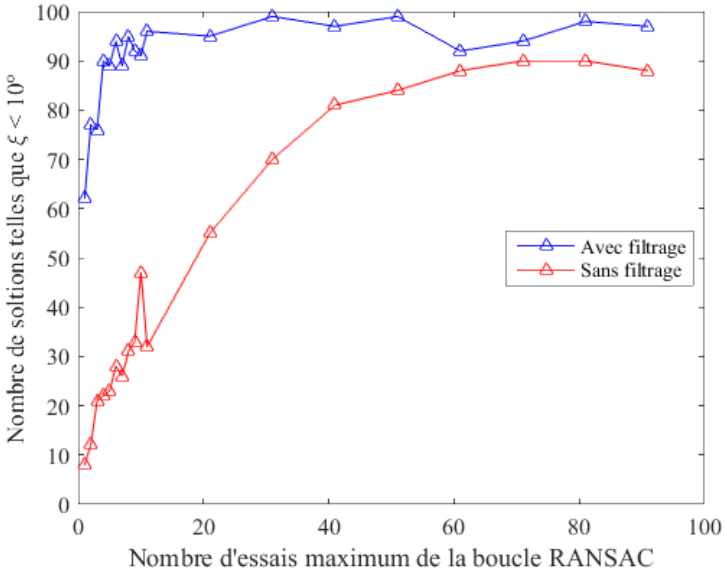


Figure 138. Nombre d'essais de la boucle *RANSAC* nécessaires à l'obtention d'une solution convenable avec (courbe rouge) et sans (courbe bleue) les étapes de filtrage des outliers.

des *outliers* résiduels. A chaque itération nous spécifions à *RANSAC* de sélectionner aléatoirement quatre correspondances. Nous allons tester le paramètre qui fixe le nombre maximum d'itération de la boucle *RANSAC*. A chaque valeur de ce paramètre, nous effectuons 100 estimations de la pose avec et sans filtrage. Nous comptons le nombre de solutions dont telles que $\xi < 10°$ (cf. section 4.2.3 équation (4.5)), c'est-à-dire, celles dont les erreurs sur les estimations des angles *yaw*, *pitch* et *roll* sont toutes les trois inférieures à 10° (cf. Figure 138). On constate qu'avec les étapes de filtrage, en moins d'une vingtaine d'itérations dans la boucle *RANSAC*, 90% des solutions sont telles que $\xi < 10°$. Si l'on se prive de ces étapes de filtrage, il faut attendre près d'une centaine d'itérations.

Nous avons également tenté d'estimer la pose avec l'algorithme *REPnP* (c'est-à-dire l'algorithme qui fonctionne sans l'aide d'une méthode *RANSAC* pour le rejet des *outliers*) dans des conditions favorables,

c'est-à-dire lorsque tous les *outliers* ont été correctement filtrés. Le résultat n'est pas satisfaisant. Ceci nous indique que l'algorithme *REPPnP* est gêné par le bruit de mise en correspondance spécifique à notre application. Nous expliquons cela par un manque de texture en imagerie thermique dans le cas où l'objet d'intérêt est un visage.

Nous avons montré jusqu'ici qu'il était possible de mettre en correspondances des points 3D-2D et d'en déduire une estimation de la pose. De plus, nous avons montré qu'une caméra étalonnée permet de réduire le temps de calcul grâce au filtrage des *outliers* basé sur le niveau continu moyen du voisinage local. Nous avons notamment testé un seuil de 320 *Adu* (sur 16 bits) et utilisé une fenêtre de voisinage local en forme d'étoile (*SWF*).

Une seule image de synthèse a été utilisée jusqu'alors. Nous allons maintenant en utiliser un grand nombre et choisir celle qui engendre la meilleure solution.

5.4.3. Choix de la meilleure solution au problème *PnP* et taille de la base d'images de synthèse

Pour une image questionnée, nous allons effectuer autant de groupes de correspondances 3D-2D que nous avons d'images de synthèse dans la base. Cette recherche exhaustive permet de ne pas utiliser l'information temporelle. Si *M* est le nombre d'images de synthèse, nous allons résoudre *M* fois le problème *PnP*. L'étape suivante consiste à rechercher la meilleure solution en se basant sur un critère d'évaluation. Nous faisons l'hypothèse que la meilleure solution minimise l'erreur de reprojection et implique un grand nombre de correspondances. Ainsi pour chaque image de synthèse, nous avons développé notre propore critère $\boldsymbol{\mathcal{C}}$ suivant :

$$\boldsymbol{\mathcal{C}}(m) = \frac{1}{N_m} + \lambda . e_m \tag{5.30}$$

L'indice $m \in \{1, \dots, M\}$ permet d'identifier l'image de synthèse parmi les *M* disponibles, N_m correspond au nombre de correspondances de la $m^{ème}$ image de synthèse et e_m l'erreur de reprojection engendrée par la $m^{ème}$ solution. λ est un poids empiriquement fixé à 0.01. La Figure 139 illustre les mises en coresspondances entre l'image réelle et les images de synthèse. Pour chacune des m images de synthèse, nous relevons N_m et e_m pour calculer $\boldsymbol{\mathcal{C}}(m)$.

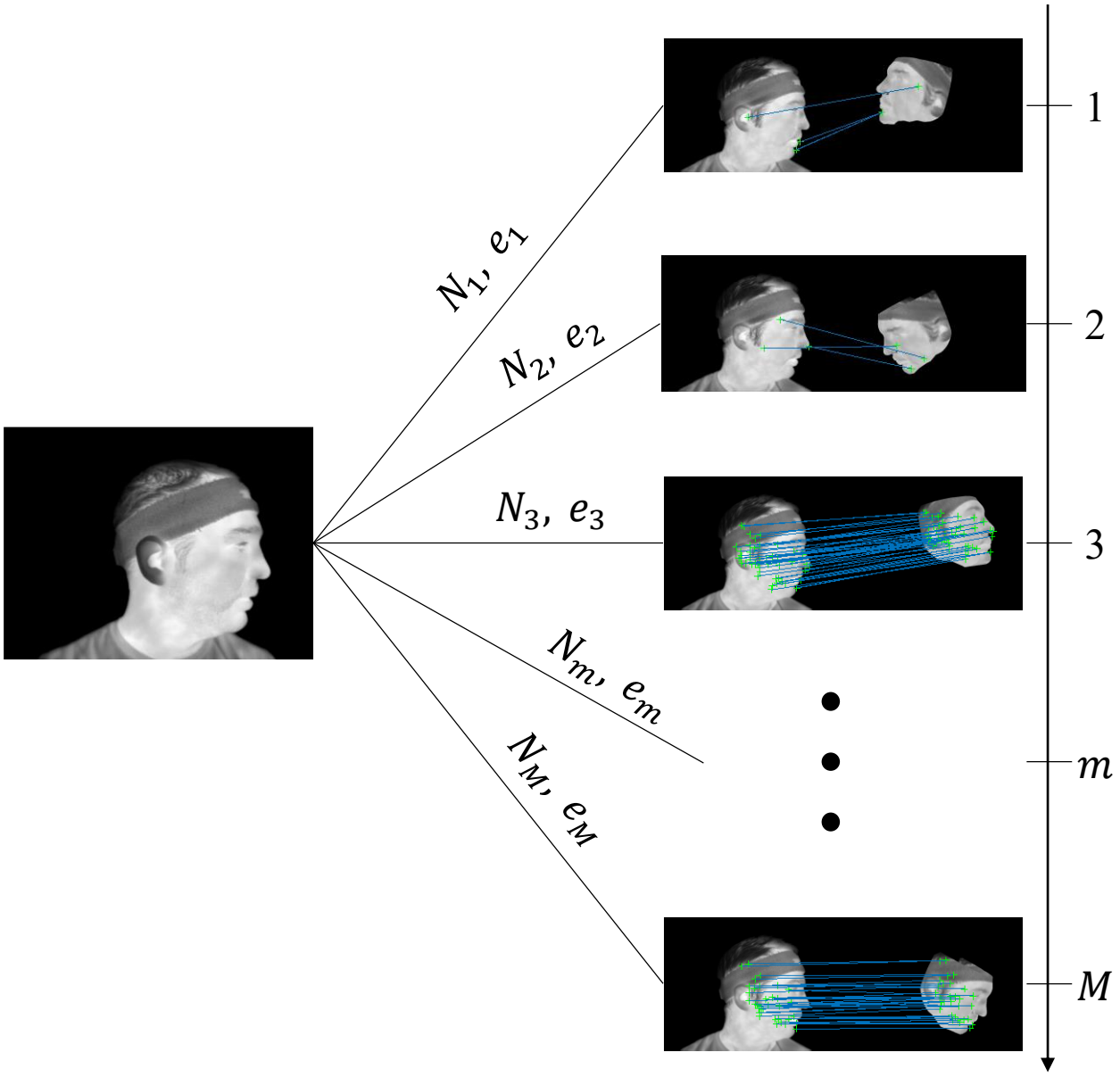


Figure 139. Représentation des mises en correspondance 3D-2D entre une image réelle et les images de synthèse de la base. Pour chacune des m images de synthèse, nous relevons N_m et e_m pour calculer $c(m)$.

L'influence du nombre d'images de synthèse de la base sur la précision de l'estimation de la pose peut être exprimée ainsi : un échantillonnage fin des angles *pitch* et *yaw* (c'est-à-dire un grand nombre de modèles) engendre un nombre de solutions précises plus important qu'un échantillonnage plus grossier (c'est-à-dire un petit nombre de modèles). Pour le voir nous avons testé et comparé trois bases d'images de synthèse. La première contient 225 modèles (*pitch* and *yaw* vont respectivement de $-20°$ à $+20°$ et de $-60°$ à $+60°$ par pas de 5°). La seconde contient 65 modèles (*pitch* and *yaw* vont respectivement de $-20°$ to $+20°$ et de $-60°$ à $+60°$ par pas de 10°). La troisième contient 45 modèles (*pitch* va de $-20°$ à $+20°$ par pas de 10° et *yaw* va de $-60°$ to $+60°$ par pas de 15°). Nous estimons dix fois de suite l'orientation du visage sur une image réelle (nous estimons plusieurs fois la pose car notre algorithme intègre la technique

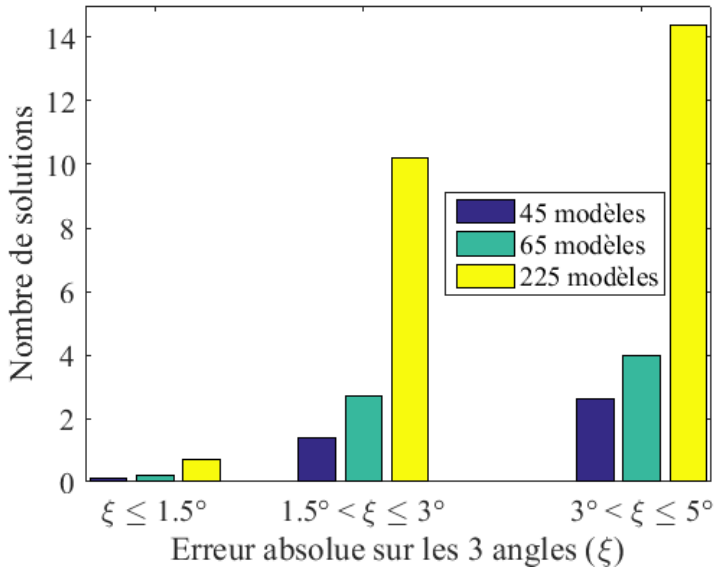


Figure 140. Influence du nombre de modèles sur la précision d'estimation d'orientation du visage.

RANSAC, donc le résultat n'est pas toujours identique pour une même image). Nous utilisons le maillage 3D précis (*Artec3D*), le filtrage *SWF* ainsi que le filtrage portant sur la norme et la direction des vecteurs de mise en correspondance. Enfin, nous comptons le nombre de solutions qui engendrent les erreurs suivantes :

- $\xi \leq 1.5°$, c'est-à-dire une erreur absolue inférieure à 1.5° sur les trois angles,
- $1.5° < \xi \leq 3°$, c'est-à-dire une erreur absolue comprise entre 1.5 et 3°,
- $3° < \xi \leq 5°$, c'est-à-dire une erreur absolue comprise entre 3 et 5°.

Les résultats sont illustrés sous forme d'un histogramme (cf. Figure 140). Si l'on souhaite améliorer le temps de calcul tout en conservant une précision inférieure à 5°, ce résultat montre qu'il est potentiellement possible d'utiliser un échantillonnage plus grossier. Or en pratique l'équation (5.30) ne permet pas toujours de choisir la solution qui minimise l'erreur sur les orientations. Ainsi, pour diminuer le temps de calcul, nous préconisons plutôt d'opter pour une autre stratégie qui conserve un maillage serré autour de la pose potentielle en utilisant une recherche multi-échelle (on parle de l'échelle d'échantillonnage en *pitch* et *yaw)*.

5.5. Résultats

Une seule vidéo est traitée dans cette section. Elle comporte 295 images numérotées de 82 à 377. Les images sont corrigées du bruit spatial grâce à une *NUC 2 points*. La table de gain est considérée indépendante de la température du détecteur. La dérive de l'offset en fonction de cette température, est compensée grâce à la méthode basée sur l'étalonnage en enceinte climatique (cf. Chapitre 2 section 2.5.4). Nous pouvons et nous utilisons le filtrage de *outliers SWF* car les images de synthèse sont créées à partir d'images réelles acquises dans les même conditions de température : il n'y a donc pas besoin d'étalonnage radiométrique.

La pose a été estimée toutes les 5 images grâce à l'algorithme illustré sur la Figure 135. Nous avons utilisé l'algorithme *SIFT* pour la détection et la mise en correspondance. La base de modèles (c'est-à-dire les images de synthèses) comporte 225 éléments comme évoqué précédemment ($yaw \in \{-60°, -55°, \dots, +60°\}$ et $pitch \in \{-20°, -15°, \dots, +20°\}$). Lorsque nous précisons que nous filtrons les *outliers*, nous faisons référence au filtrage *SWF* et à celui basé sur la norme et la direction des vecteurs de mise en correspondance. Pour la résolution du problème *PnP*, l'algorithme *RPnP* est intégré dans un schéma *RANSAC*. Dans celui-ci, quatre correspondances 3D-2D sont sélectionnées au hasard et utilisées pour estimer la pose à chaque itération. Nous avons limité le nombre d'itérations à 25. La meilleure pose parmi ces 25 permet de définir des *outliers*, c'est-à-dire les correspondances 3D-2D dont l'erreur de reprojection est supérieure à 10 pixels et des *inliers*. Une dernière estimation de la pose est réalisée en utilisant uniquement les *inliers* (dans cette dernière étape on utilise donc $n \geq 4$ correspondances 3D-2D).

Comme présenté dans la section 4.2, nous évaluons quantitativement les estimations de la pose en comparant les orientations *yaw*, *pitch* et *roll* estimées, aux orientations mesurées par une centrale inertielle portée par le participant. La Figure 141 présente les mesures lorsqu'un maillage 3D précis est utilisé (Artec3D). La Figure 142 présente les mesures lorsqu'un maillage 3D grossier est utilisé (ellipsoïde).

Pour chaque angle (*yaw*, *pitch* et *roll*), nous évaluons l'erreur moyenne (*EM*) et l'écart type de l'erreur (*S*) sur les 59 estimations (la vidéo comporte 295 images et on réalise une estimation toutes les 5 images). Le Tableau 12 répertorie les résultats lorsqu'un maillage 3D précis est utilisé (Artec3D). Le Tableau 13 répertorie les résultats lorsqu'un maillage 3D grossier est utilisée (ellipsoïde). Avec chaque maillage, nous avons estimé la pose sans les filtrages d'*outliers* et avec le filtrage *SWF* et le filtrage sur la norme et l'argument du vecteur d'appariement.

Tableau 12. Erreurs d'estimation avec un maillage précis (Artec3D).

	Résultats des estimations sur les rotations ($\boldsymbol{EM \pm S}$)		
Filtrages *outliers*	**Erreur sur *pitch***	**Erreur sur *yaw***	**Erreur sur *roll***
Sans	-1.2° ± 9,0°	3.0° ± 4.9°	1.7° ± 7.1°
Avec	-3.1° ± 5,7°	3.2° ± 3.6°	1.1° ± 3.8°

Tableau 13. Erreurs d'estimation avec un maillage grossier (ellipsoïdal).

	Résultats des estimations sur les rotations ($\boldsymbol{EM \pm S}$)		
Filtrages *outliers*	**Erreur sur *pitch***	**Erreur sur *yaw***	**Erreur sur *roll***
Sans	-1.7° ± 10.3°	3.7° ± 8.0°	-0.7° ± 9.3°
Avec	-3.4° ± 6.0°	2.5° ± 5.0°	0.7° ± 4.1°

Les Tableau 12 et Tableau 13, montrent une erreur moyenne sur l'angle *pitch* plus faible sans les filtrages des *outliers* qu'avec. Ceci est contrintuitif car les filtrages des *outliers* permettent, en théorie, d'améliorer l'estimation. Pour comprendre ce chiffre, regardons les estimations de l'angle *pitch*. La Figure 141 répertorie les estimations obtenues avec le maillage précis et la Figure 142, celles obtenues avec le

maillage ellipsoïdal. Le premier graphique de chaque figure concerne spécifiquement l'angle *pitch*. On constate que les estimations sans filtrage des *outliers* (courbes en pointillées bleues) sont affectées par un bruit blanc de grande amplitude.

D'après les Tableau 12 et Tableau 13, il apparait, comme on pouvait s'y attendre, que l'estimation de l'orientation est meilleure lorsqu'un maillage précis est utilisé. De plus lorsque le nombre d'itérations maximum de la méthode RANSAC est fixé à 25, il apparait que l'utilisation des briques de filtrages (*SWF* et norme et direction des vecteurs d'appariements) améliore l'estimation de l'orientation.

Pour visualiser qualitativement les résultats, nous avons projeté le maillage 3D grâce aux paramètres intrinsèques et à l'estimation de la pose. Sur la Figure 143 le maillage précis est utilisé pour trouver la 3D dans les correspondances 3D-2D. C'est pourquoi c'est ce maillage qui est projeté pour visualiser les résultats. De la même manière, sur la Figure 144, le maillage ellipsoïdal a été utilisé, il est donc projeté.

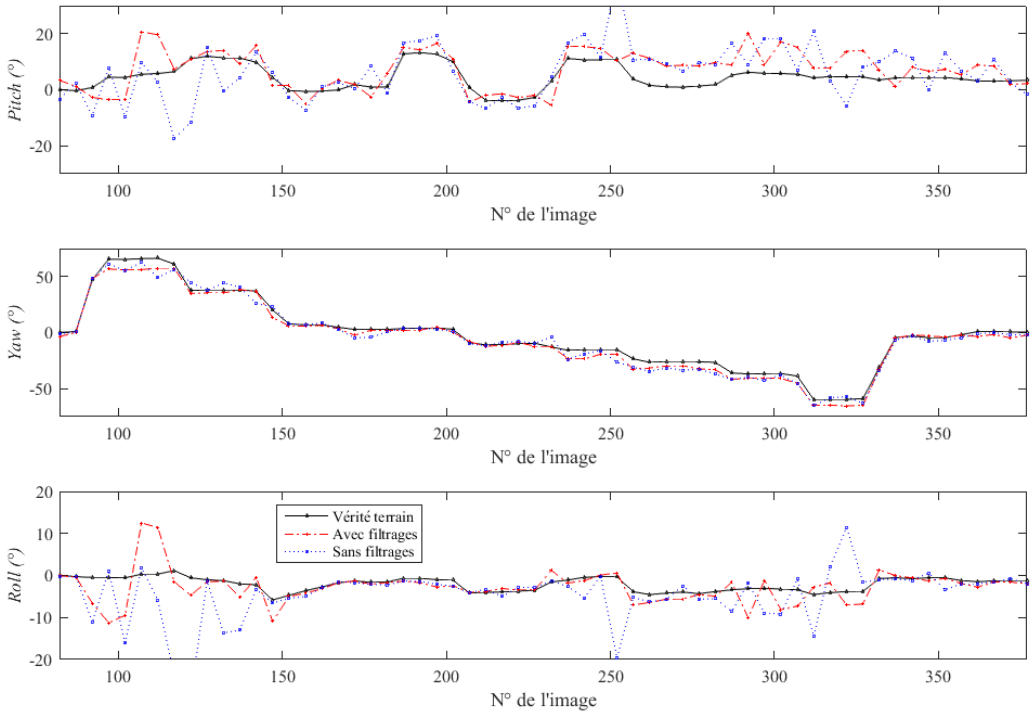


Figure 141. Résultats obtenus avec un maillage précis (Artec3D).

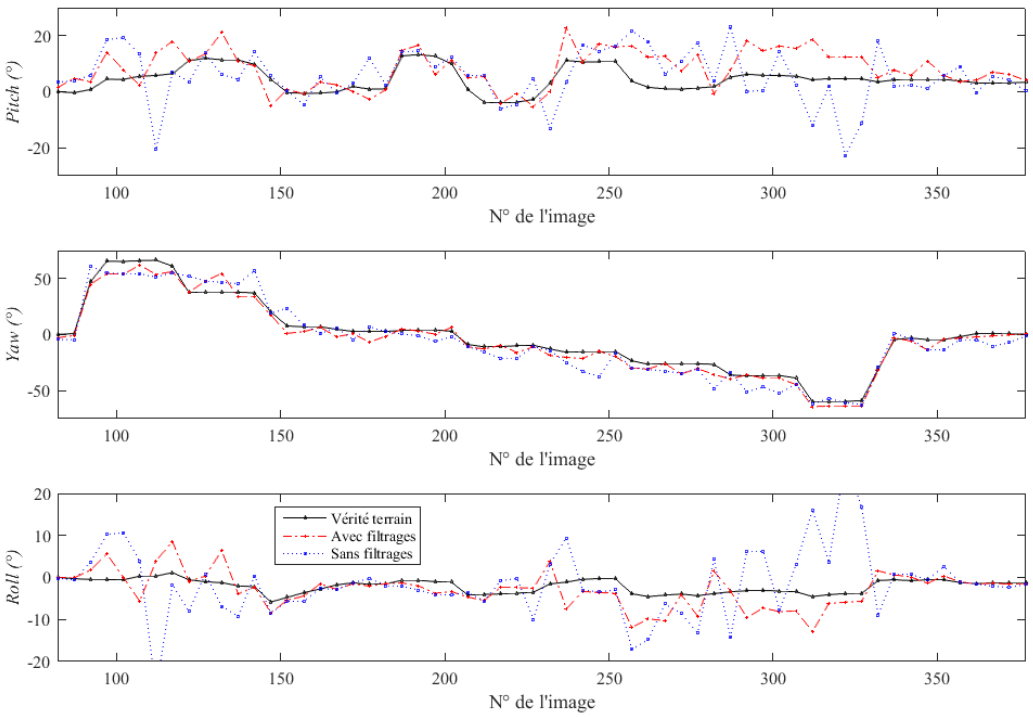


Figure 142. Résultats obtenus avec un maillage grossier (ellipsoïdal).

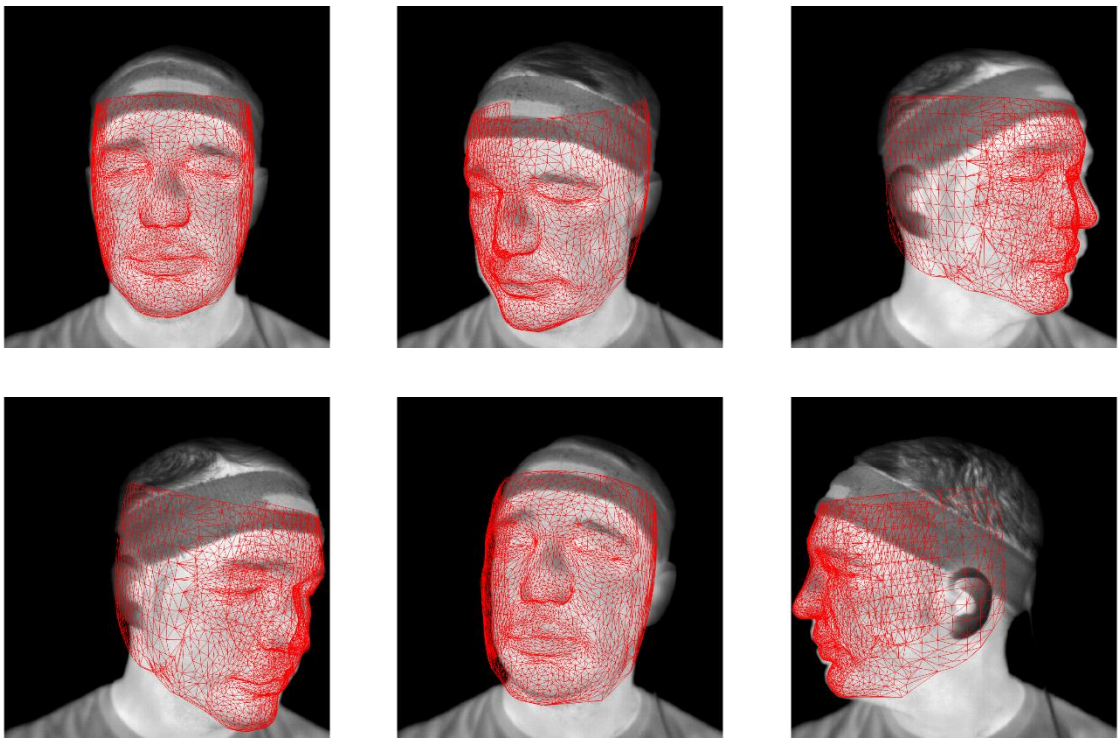


Figure 143. Projection du maillage 3D grâce à l'estimation de la pose. Le maillage utilisé est précis (Artec3D).

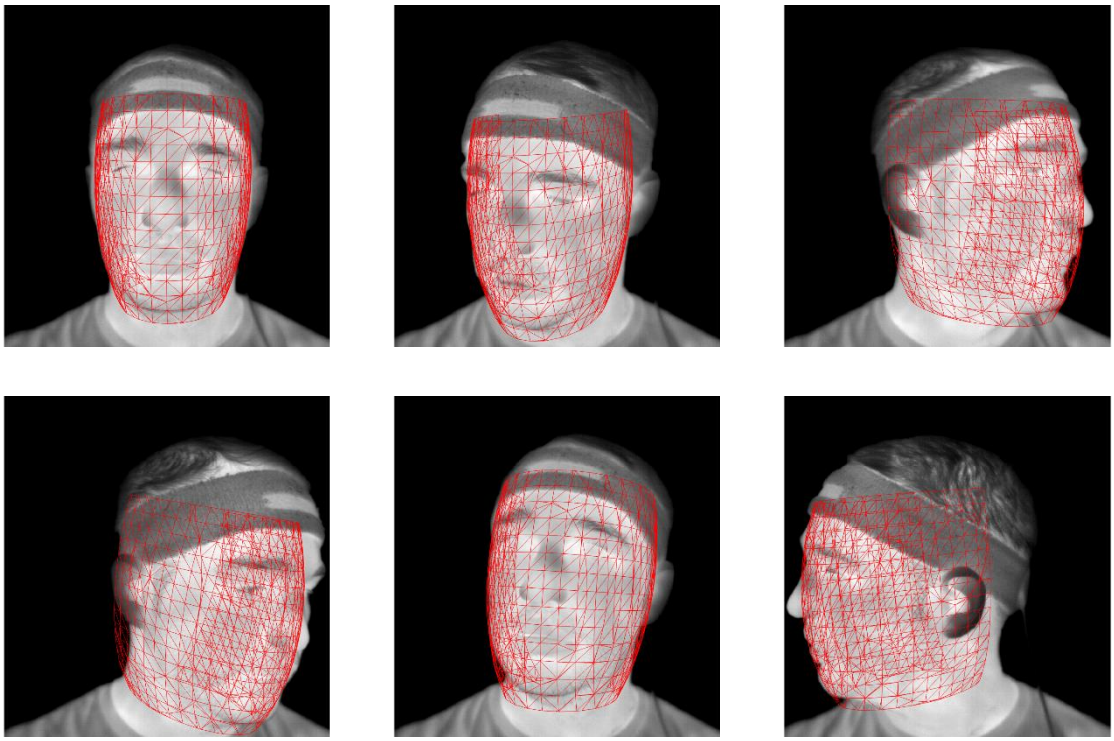


Figure 144. Projection du maillage 3D grâce à l'estimation de la pose. Le maillage utilisé est grossier (ellipsoïde).

Chapitre 6. Estimation de la pose à partir d'un capteur bas coût simulé

6.1. Introduction

Il est intéressant d'évaluer les résolutions thermique et spatiale minimales permettant d'adresser l'application d'estimation de la pose du visage afin de ne pas surspécifier le système et ainsi espérer un système bas coût. Nous disposons d'une base d'images acquises avec un capteur au format 640×480 au pas de 17 µm associé à une optique ouverte à f/0.8 avec un champ horizontal de 53°. La *NETD* de ce système a été évaluée à 60 mK. Nous cherchons à savoir si un capteur moins cher permettrait d'atteindre des performances équivalentes en termes de précision d'estimation de la pose. Pour cela nous allons dégrader en format de détecteur et en *NETD* cette base d'images et appliquer les algorithmes d'estimation de la pose vus précédemment.

6.2. Réduction du nombre de pixels et du pas pixel

Réduire le nombre de pixels est un levier permettant de réduire le coût du système global. Nous allons donc simuler l'image qui aurait été obtenue avec un détecteur de format plus petit à partir des images acquises avec la caméra *GOBI 640 CL*. Le détecteur que nous simulons, nous l'appelons « capteur bas coût » en comparaison avec le coût d'un capteur 640×480 de la caméra *Gobi*. Nous considérons que les pixels sont carrés. Pour la caméra bas coût, nous notons :

- $t_{bas\ coût}$ la taille d'un côté de la surface photosensible du pixel,
- $p_{bas\ coût}$ le pas pixel,
- $f_{bas\ coût}$ la distance focale de ce capteur,
- $N_{bas\ coût}$ le nombre de pixels dans la largeur,
- $FOV_{bas\ coût}$ le champ de vue.

De même, pour la caméra *Gobi* nous notons :

- t_{Gobi} la taille d'un côté de la surface photosensible du pixel,
- p_{Gobi} le pas pixel,
- f_{Gobi} la distance focale de ce capteur,
- N_{Gobi} le nombre de pixels dans la largeur.
- FOV_{Gobi} le champ de vue.

Le champ et l'ouverture sont considérés équivalents entre la caméra *Gobi 640 CL* et le capteur bas coût $FOV_{bas\ coût} = FOV_{Gobi}$ dans cette section.

Notre simulation prend en compte uniquement le filtrage du pixel et pas le filtrage de l'optique. Nous allons filtrer l'image pour simuler le filtrage d'un pixel équivalent plus gros que le pixel réel de la caméra *Gobi*. Attention, ce filtrage va améliorer le bruit temporel, et donc la *NETD* du système. Nous traiterons cela dans un deuxième temps dans la section suivante.

Dans le domaine spatial, le filtrage d'un pixel est modélisable par une convolution avec la fonction rectangle $Rect(x/t_x)$ le long de l'axe horizontal et par la fonction $Rect(y/t_y)$ le long de l'axe vertical. La fonction rectangle est définie comme suit :

$$Rect\left(\frac{x}{t_x}\right) = \begin{cases} 1 \; si \; |x| \leq \frac{1}{2} t_x \\ 0 \; sinon \end{cases} \tag{6.1}$$

La taille d'un côté de la surface photosensible du pixel est t_x selon l'axe horizontal et t_y selon l'axe vertical. Dans le cas d'un pixel de la caméra bas coût, les pixels sont carrés donc $t_x = t_y = t_{bas\ cout}$. Il en est de même avec la caméra *Gobi*. Le résultat du produit de convolution de l'image avec les fonctions $Rect(x/t_x)$ et $Rect(y/t_y)$ correspond à un filtrage passe bas inhérent à l'utilisation de pixels pour créer une image.

Dans le domaine fréquentiel, le filtrage du pixel est modélisable par la multiplication de la transformé de Fourier de l'image avec la transformé de Fourier de la fonction $Rect(x/t_x)$ qui s'exprime ainsi :

$$\mathrm{TF}\left[Rect\left(\frac{x}{t_x}\right)\right] = \mathrm{sinc}(\nu_x . t_x), \tag{6.2}$$

avec ν_x la fréquence spatiale le long de l'axe horizontal (la transformée de Fourier de la fonction $Rect(y/t_y)$ est similaire et dépend de la fréquence spatiale ν_y). La définition du sinus cardinal utilisée est $\mathrm{sinc}\, x = \frac{\sin(\pi x)}{\pi x}$.

Si l'on ne tient pas compte de la fonction de transfert de l'optique, et qu'on s'intéresse uniquement à la fonction de transfert du pixel, l'image **Y** numérisée par le capteur est modélisable par l'équation suivante :

$$\mathbf{Y} = \mathrm{TF}^{-1}\left[\mathrm{TF}[\boldsymbol{Z}] \times \mathrm{sinc}(\nu_x . t_x) \times \mathrm{sinc}(\nu_y . t_y)\right] \tag{6.3}$$

L'image idéale (sans filtrage par les pixels) est notée $\boldsymbol{Z}$.

Modifier le nombre et la taille des pixels à champ de vue constant implique de modifier la distance focale. La Figure 145 représente la position du plan image de la caméra *Gobi 640 CL* et celui du capteur bas coût. Le champ, le pas pixel et la distance focale sont liés par la relation suivante :

$$\tan\left(\frac{FOV_{bas\ coût}}{2}\right) = \frac{N_{bas\ coût} \times p_{bas\ coût}}{2 \times f_{bas\ coût}} \tag{6.4}$$

Nous pourrions écrire la même relation pour la caméra *Gobi*.

Si on fait l'hypothèse d'un facteur de remplissage de 100 % alors il est possible remplacer dans l'équation (6.4) le pas pixel $p_{bas\ coût}$ par la taille de la surface photosensible $t_{bas\ coût}$ et on obtient :

$$\tan\left(\frac{FOV_{bas\ coût}}{2}\right) = \frac{N_{bas\ coût} \times t_{bas\ coût}}{2 \times f_{bas\ coût}} \tag{6.5}$$

On pourrait faire la même chose avec la caméra *Gobi*.

Enfin, Le fait de conserver un champ constant ($FOV_{bas\ coût} = FOV_{Gobi}$) impose une contrainte sur le rapport entre les distances focales :

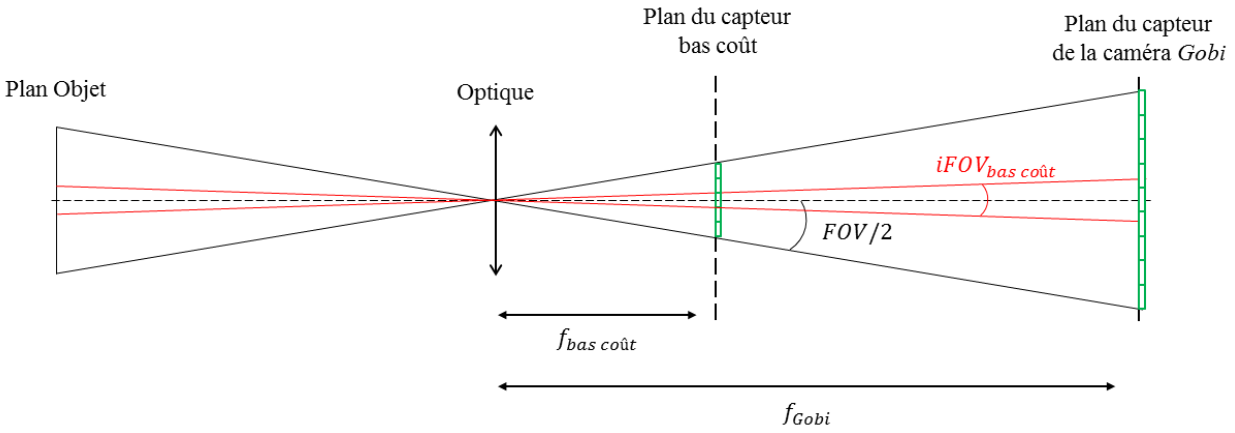


Figure 145. Illustration du changement de format (nombre et taille des pixels) à champ de vue constant

$$\frac{f_{bas\ coût}}{f_{Gobi}} = \frac{N_{bas\ coût} \times t_{bas\ coût}}{N_{Gobi} \times t_{Gobi}} \tag{6.6}$$

Si on considère uniquement le filtrage dû à la surface photosensible du pixel, une image obtenue par la caméra bas coût dans le plan du capteur bas coût est équivalente à une image obtenue dans le plan du capteur de la caméra *Gobi* avec une taille de pixel $t_{équivalent}$ liée à $t_{bas\ coût}$ comme suit :

$$t_{équivalent} = t_{bas\ coût} \frac{f_{Gobi}}{f_{bas\ coût}} \tag{6.7}$$

En injectant l'équation (4.1) dans l'équation (6.7) on obtient :

$$t_{équivalent} = \frac{N_{Gobi} \times t_{Gobi}}{N_{bas\ coût}}\ avec\ N_{bas\ coût} < N_{Gobi} \tag{6.8}$$

Le nombre de pixels de la caméra *Gobi* est $N_{Gobi} = 640$ et la taille du pixel est $t_{Gobi} = 17\ \mu m$.

$$t_{équivalent} = \frac{640 \times 17\ \mu m}{N_{bas\ coût}}\ avec\ N_{bas\ coût} < 640$$

Lorsque l'on réduit la taille du capteur, c'est-à-dire le nombre et la taille des pixels, et qu'on souhaite conserver un champ constant, cela revient à considérer que la taille équivalente $t_{équivalent}$ du pixel, dans le plan du capteur initial, est supérieure à la taille du pixel initial ($17\ \mu m$) comme cela est exprimé par l'équation (6.7).

On peut simuler le filtrage du pixel équivalent de taille $t_{équivalent}$ à partir de l'image acquise grâce à des pixels de taille t_{Gobi} en deux opérations. La première consiste à rehausser le contratse qui a été diminué par le filtrage du pixel réel de la caméra *Gobi*. Pour cela, nous divisons le transformé de Fourier de l'image réelle par $\mathrm{sinc}(\nu_x . t_{Gobi})$ et $\mathrm{sinc}(\nu_y . t_{Gobi})$. Le seconde opérations, consiste à diminuer le contraste pour simuler le filtrage du pixel équivalent. Pour cela, dans l'espace de Fourier, nous multliplions par $\mathrm{sinc}(\nu_x . t_{équivalent})$ et $\mathrm{sinc}(\nu_y . t_{équivalent})$. Ces deux opérations se résument finalement en une seule équation :

$$\mathrm{TF}[\boldsymbol{Y}_{bas\ coût}] = \mathrm{TF}[\boldsymbol{Y}] \times \frac{\mathrm{sinc}(\nu_x . t_{équivalent})}{\mathrm{sinc}(\nu_x . t_{Gobi})} \times \frac{\mathrm{sinc}(\nu_y . t_{équivalent})}{\mathrm{sinc}(\nu_y . t_{Gobi})} \qquad (6.9)$$

L'espace fréquentiel est échantillonné au pas de fréquence F_e/N_{Gobi} le long l'axe horizontal et au pas de fréquence $F_e/(\frac{3}{4} * N_{Gobi})$ le long de l'axe vertical avec $F_e = 1/t_{Gobi}$.

Remarque : la taille du pixel équivalent $t_{équivalent}$ ne dépend pas de la taille $t_{bas\ coût}$ du pixel de la caméra bas coût comme cela a été exprimé dans l'équation (6.8). Donc l'image simulée, comme le montre l'équation (6.9), ne tient pas compte non plus de $t_{bas\ coût}$. En effet, dans notre simulation, nous considérons que l'optique ne limite pas la résolution spatiale du système. En pratique, un pixel plus petit engendre des complications dans le design optique de l'objectif. En effet, le diamètre de la réponse percussionnelle de l'optique doit diminuer pour ne pas devenir l'élément limitant en termes de résolution spatiale. Un objectif parfait, c'est-à-dire sans aberration optique, possède un pouvoir de résolution spatiale qui est limité par la diffraction. Rappelons que la limite de diffraction optique est proportionnelle au nombre d'ouverture N, d'après l'expression du diamètre de la tâche d'Airy : $\phi_{Airy} \approx 1{,}22.\lambda.N$. Une stratégie possible consiste à augmenter l'ouverture, c'est-à-dire diminuer l'ouverture numérique N, de la caméra pour garantir que l'optique ne devienne pas l'élément limitant lorsque l'on réduit la taille des pixels. Cette stratégie n'est pas envisageable en imagerie thermique car les optiques ont la particularité d'être très ouvertes pour garantir une bonne sensibilité thermique. La diminution de la taille du pixel est donc inévitablement accompagnée d'une perte de contraste liée au fait que l'optique devient l'élément limitant du système en termes de résolution spatiale.

De plus, un petit pixel absorbe moins de photons, il est donc nécessaire de compenser cela par un circuit de lecture optimisé ou une plus grande ouverture de l'optique.

Pour illustrer cette opération, modélisons le filtrage des pixels carrés d'une matrice qui comporte 80×60. Le pixel équivalent dans le plan de la caméra *Gobi* vaut $t_{équivalent} = 136\ \mu m$. Le résultat du filtrage par le pixel équivalent dans l'espace de Fourier est illustré sur la Figure 146. L'image réelle a) acquise par la caméra *Gobi* est filtrée grâce au filtre représenté en b). Le résultat obtenu en c) est une image lissée. On peut constater sur les agrandissements, au niveau du sourcil, que les hautes fréquences sont perdues. Après le filtrage, il est nécessaire d'échantillonner correctement l'image. nous prenons simplement un pixel sur huit (car 640/80=8) de l'image filtrée, on obtient l'image, d) de la Figure 146, au format 80×60. Cette image a été agrandie pour l'affichage. L'image à taille réelle est située en haut dans l'encadré.

Remarquons que l'équation (6.9) a pour effet de lisser l'image. Ainsi la valeur d'un pixel va être remplacée par une moyenne pondérée de son voisinage local. Le bruit temporel, c'est-à-dire le bruit responsable de la *NETD*, va donc être diminué. L'adaptation de la *NETD* à la simulation de la réduction du format du détecteur est l'objet de la section suivante.

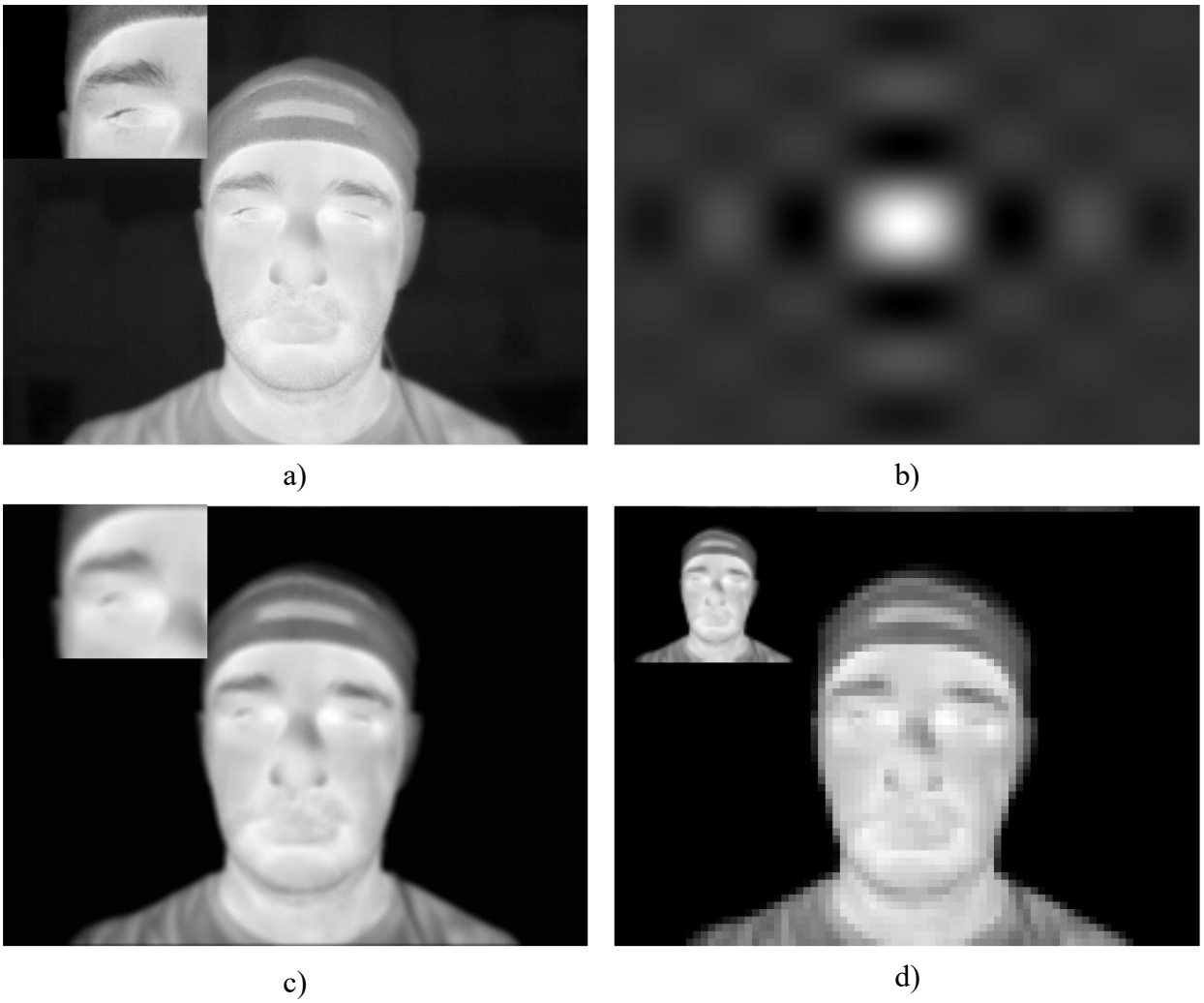
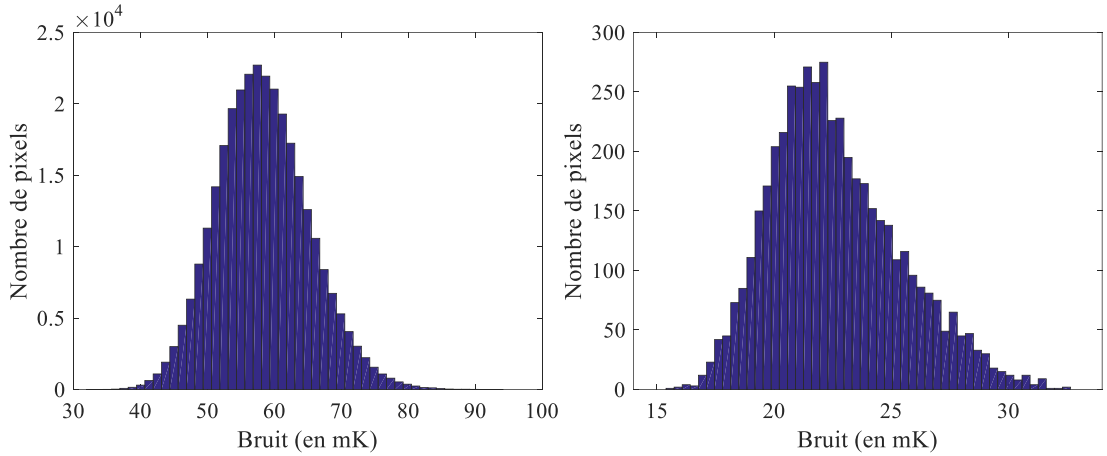


Figure 146. Filtrage de l'image acquise par la caméra *Gobi* pour simuler le filtrage pixel d'une caméra de 80x60 pixels. a) Image acquise avec la caméra *Gobi* (au format 640×480 au pas de 17 µm). Un agrandissement de l'œil gauche est représenté en haut à gauche. b) Fonction de filtrage dans l'espace de Fourier. c) Image ayant subi le filtrage du pixel équivalent. Un agrandissement de l'œil gauche est représenté en haut à gauche. d) Image filtrée et échantillonnée au format 80×60. Cette image fait la même taille que l'image 640×480 car, pour l'affichage, nous avons recopié huit fois chaque pixel de l'image au format 80×60. L'image à taille réelle est représentée en haut à gauche.

6.3. Prise en compte de la *NETD*

Lorsqu'on réduit le format du détecteur de la caméra *Gobi* en filtrant les pixels, on améliore la *NETD*. Nous pouvons nous rendre compte de l'amélioration de la *NETD* en appliquant l'opération de filtrage du pixel à des images du corps noir et en calculant à nouveau (comme dans le chapitre 2) le bruit temporel en *Adu*, la *responsivité* en *Adu*/K et enfin la *NETD*. La *NETD* de l'image originale de la caméra *Gobi* et la *NETD* de l'image l'image simulée au format 80×60 pixels sont représentées sur la Figure 147. On constata effectivement une amélioration de la *NETD* après le filtrage du pixel équivalent.



a) b)

Figure 147. Bruit temporel évalué grâce aux images du corps noir à 30°C et 40°C après filtrage du pixel équivalent. La taille du pixel équivalent (qui vaut 136 µm) a été calculée pour passer du format 640×480 au pas de 17 µm à un format de 80×60 pixels tout en gardant le champ de vue constant. a) *NETD* de l'image originale *Gobi*. b) *NETD* de l'image simulée au format 80×60 pixels.

La médiane de la *NETD* mesurée à partir des images originales acquises avec la caméra *Gobi* vaut 58 mK (la moyenne vaut 58.36 mK). La médiane de l'histogramme des valeurs de la *NETD* vaut 22.2 mK (la moyenne vaut 22.6 mK). Nous souhaitons pouvoir revenir à une *NETD* d'environ 60 mK car une *NETD* autour de 22 mK implique une augmentation du prix du capteur. Nous souhaitons donc simuler une augmentation du bruit temporel. De plus, nous aimerions pouvoir tester d'autres niveaux de *NETD* pour obtenir des images représentatives des capteurs à bas coût. Nous simulerons plusieurs niveaux de *NETD* :

- Dans un premier temps, 60 mK pour retrouver la *NETD* de la caméra *Gobi* et ainsi pouvoir analyser uniquement l'impact du format de la matrice sur la qualité de l'estimation de la pose.
- Ensuite 100, 120, 150, 200, 250, … mK pour spécifier notre besoin en *NETD* pour l'application précise d'estimation de la pose du visage.

Afin d'augmenter le niveau de *NETD* artificiellement, nous allons nous appuyer sur un modèle expliquant la manière dont différentes sources de bruit contribuent à la *NETD* finale. Nous ne nous intéressons pas à la nature du bruit que nous ajoutons, nous sommes simplement intéressés par l'augmentation du niveau final de la *NETD*. Nous considérons que la réponse du capteur est entachée d'un bruit en tension représenté par son écart type $\sigma_{temporel}$ (en V ou en *Adu*), qu'il est possible d'exprimer par

son équivalence en température (la *NETD*) grâce à la réponse du capteur (en *Adu*/K). Le bruit temporel total peut être exprimé sous la forme d'une somme quadratique des différentes sources [80]. Par exemple, si on considère trois sources de bruit temporel représentées par σ_a, σ_b et σ_c, le bruit temporel total $\sigma_{temporel}$ s'exprime ainsi :

$$\sigma_{temporel}^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2 \tag{6.10}$$

Pour simuler une augmentation du bruit sur les images obtenues après le filtrage du pixel équivalent, nous proposons d'ajouter un bruit d'écart type $\sigma_{ajouté}$ (en V ou en *Adu*) au bruit déjà présent sur l'image obtenue après filtrage du pixel équivalent. Nous évaluons la quantité $\sigma_{ajouté}$ comme suit :

$$\sigma_{ajouté}^2 = \sigma_{ciblé}^2 - \sigma_{lissé}^2 \tag{6.11}$$

Le terme $\sigma_{ciblé}$ représente l'écart type du bruit temporel ciblé (en V ou en *Adu*) et le terme $\sigma_{lissé}$ représente l'écart type du bruit temporel après avoir appliqué le filtrage du pixel bas coût.

La spécification technique donnée par les fournisseurs pour évoquer le bruit temporel est souvent la *NETD*. Sur une petite plage de température de la scène et pour une température fixe du capteur, la *responsivité* du détecteur $\mathfrak{R}_{Adu/K}$ en *Adu*/K peut être considérée comme linéaire. Donc lorsque nous visons une $NETD_{ciblée}$ donnée, nous visons un bruit temporel d'écat type $\sigma_{ciblé}$ qui peut être évalué approximativement par :

$$\sigma_{ciblée} = NETD_{ciblée} \times \langle \mathfrak{R}_{Adu/K} \rangle \tag{6.12}$$

La *responsivité* $\mathfrak{R}_{Adu/K}$ de l'équation ci-dessus est obtenue après filtrage du pixel équivalent. Elle est évaluée grâce à deux températures du corps noir qui valent 30°C et 40°C et à température fixe de la caméra. Nous avons pris la moyenne spatiale de la matrice des *responsivités* $\langle \mathfrak{R}_{Adu/K} \rangle$.

Par exemple, faisons l'application numérique lorsque nous visons $NETD_{ciblée} = 100\,mK$ à une température capteur de 38°C pour un format capteur de 160×120 au pas de 12 µm :

- la moyenne spatiale de la *responsivité* $\langle \mathfrak{R}_{Adu/K} \rangle$ mesurée à partir d'images (après le filtrage du pixel équivalent) du corps noir à 30°C et à 40°C vaut 172.1 *Adu*/K.
- On en déduit l'écart type du bruit temporel ciblée grâce à l'équation (6.12) et on obtient $\sigma_{ciblée} = 17.2\,Adu$.
- La médiane de l'écart type du bruit temporel après le filtrage du pixel équivalent vaut $\sigma_{lissé} = 4.57\,Adu$ (cf. Figure 147 a).
- Grâce à l'équation (6.11), on évalue la médiane de l'écart type à ajouter à l'image filtrée par le pixel équivalent et on obtient $\sigma_{ajouté} = 16.58\,Adu$.

Sur chaque pixel de l'image ayant subi le filtrage du pixel équivalent, nous allons donc ajouter un bruit représenté par une variable aléatoire qui suit la loi normale $\mathcal{N}(0, \sigma_{ajouté})$. Ce bruit est blanc car sa densité spectrale de puissance est identique pour toutes les fréquences spatiales. Le résultat de l'ajout de ce bruit est représenté sur la Figure 148. Nous avons représenté plusieurs images en conservant toujours le format initial de 640×480 pixels tout en simulant différents niveaux de *NETD*, 60, 150, 200, 250, 300 et 350 mK.

a) b)

c) d)

e) f)

Figure 148. Ajout d'un bruit simulé sur l'image initiale de la caméra *Gobi* pour atteindre des niveaux de *NETD* supérieurs. a) Image initiale *NETD* = 60 mK, b) *NETD* = 150 mK, c) *NETD* = 200 mK, d) *NETD* = 250 mK, e) *NETD* = 300 mK, d) *NETD* = 350 mK.

A titre d'illustration, nous avons représenté sur la Figure 149 quatre formats 640×480, 320×240, 160×120 et 80×60. Nous avons simulé une *NETD* de 200 mK sur ces quatres formats.

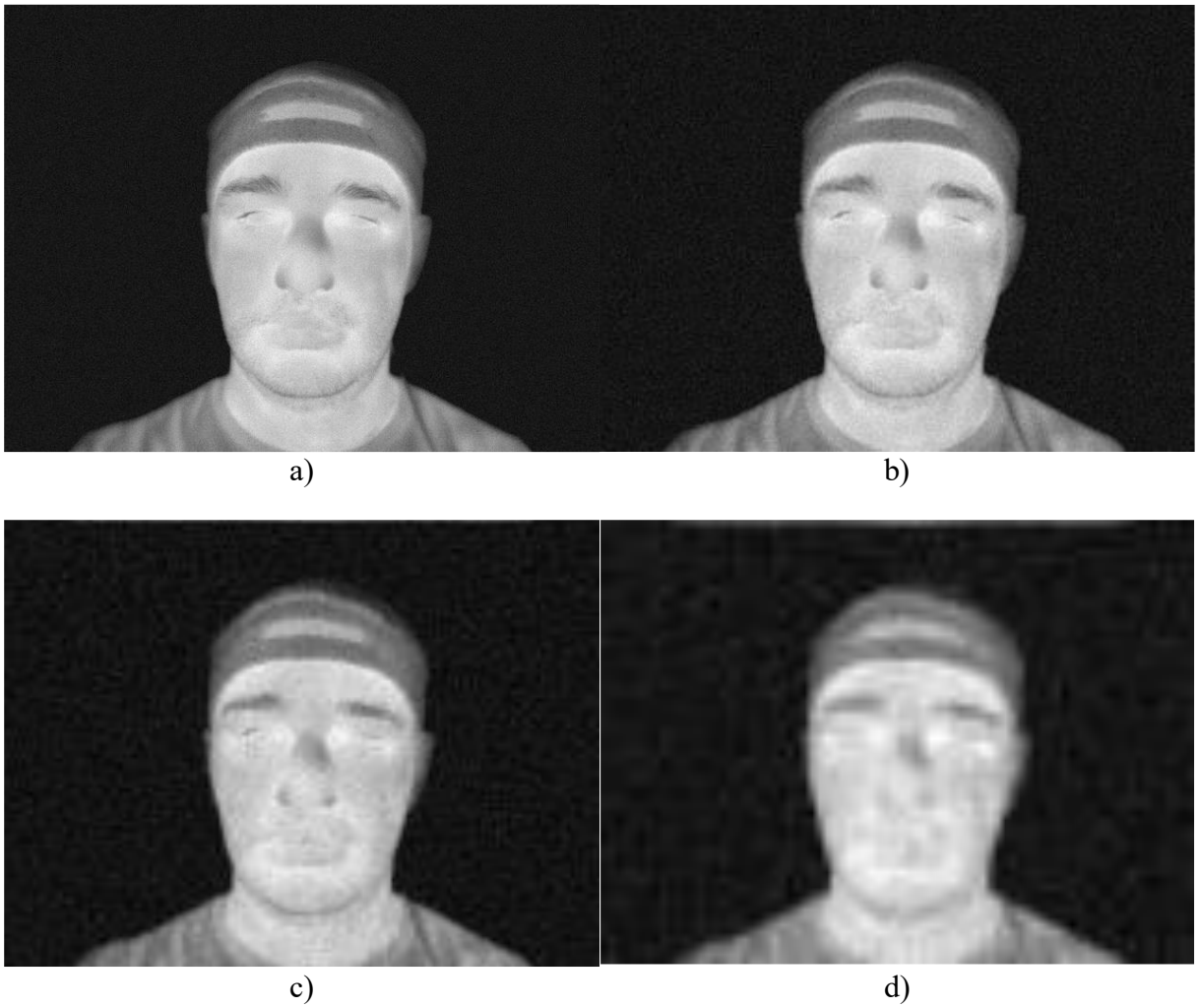


Figure 149.Quelques images issues du modèle capteur. a) Image au format 640×480 pixels avec une *NETD* de 200 mK. b) Image au format 320×240 pixels avec une *NETD* de 200 mK. c) Image au format 160×120 pixels avec une *NETD* de 200 mK. d) Image au format 80×60 pixels avec une *NETD* de 200 mK.

Nous allons tester plusieurs configurations pour lesquelles le format de la matrice en pixels pourra être 80×60, 160×120, 320×240 et 640×480 pixels. Les niveaux de *NETD* testés seront 60, 100, 120, 150, 200, 250, … mK. Toutes les configurations sont listées dans l'annexe E, où nous avons mentionné le bruit en *Adu* à ajouter pour atteindre la *NETD* visée. Un contrôle systématique de la *NETD* est effectué après avoir ajouté du bruit à des images du corps noir ayant subi le filtrage du pixel équivalent. Ces valeurs de contrôle apparaissent également dans l'annexe E. On observe que notre modèle capteur permet d'obtenir des *NETD*, contrôlées sur les images du corps noir, qui sont proches des *NETD* ciblées.

6.4. L'algorithme « local »

Pour tester l'algorithme « local » détaillé au chapitre 5, une seule vidéo de 300 images est utilisée dans laquelle le visage adopte des orientations typiques dans un contexte de conduite. Cette vidéo est acquise avec la caméra *Gobi* (format 640×480 pixels pour un champ de vue horizontal de 53° et avec une *NETD* d'environ 60 mK). Les autres formats et niveaux de *NETD* sont simulés à partir des images originales comme cela est détaillé dans les sections précédentes.

Nous faisons le choix de considérer qu'un étalonnage radiométrique a été effectué. Ainsi, il est possible de supprimer une partie des *outliers* en s'appuyant sur le niveau continu moyen d'une fenêtre de 5×5 pixels autour du point d'intérêt. Si la différence de niveau dépasse le seuil de 320 *Adu* sur une image 16 bits (ce qui représente une variation de 1.8°C à $T_C = 38°C$), la mise en correspondance est considérée comme un *outlier*. L'image initiale est codée en 16 bits mais un passage en 8 bits est réalisé car les images de synthèse sont sauvegardées en 8 bits. Nous prenons 1600 niveaux de l'image 16 bits (ce qui représente une dynamique approximative de 9.2°C à $T_C = 38°C$), puis nous étalons la dynamique et nous convertissons l'image en 8 bits. Le seuil de 320 *Adu* sur 16 bits devient un seuil de 51 niveaux en 8 bits. Nous limitons l'algorithme *RANSAC* à 25 itérations. Le maillage précis et adapté à mon visage (obtenu avec le scanner 3D de la société *Artec 3D*) a été utilisé.

Nous avons testé plusieurs niveaux de *NETD* sans modifier le format initial de la caméra *Gobi* (format 640×480 pixels au format de 17 µm). A chaque image d'une vidéo, nous relevons l'erreur maximale ξ parmi les trois angles *yaw*, *pitch* et *roll* commise. Nous rappelons ci-dessous l'expression de ξ établie au Chapitre 4 (cf. équation (4.5)) :

$$\xi(i) = \left|\max\left(e_{yaw}(i), e_{pitch}(i), e_{roll}(i)\right)\right|$$

L'indice i permet de repérer l'image dans la vidéo ($i \in [1,300]$). Les termes $e_{yaw}(i)$, $e_{pitch}(i)$, et $e_{roll}(i)$ indiquent les erreurs commises sur les angles *yaw*, *pitch* et *roll* à la $i^{ième}$ image.

Nous avons estimé ξ sur une vidéo de 300 images et pour plusieurs niveaux de *NETD* simulés. A chaque niveau de *NETD* nous calculons le pourcentage d'images qui respecte $\xi < 8°$ (cf. Figure 150). On s'aperçoit logiquement que lorsque la *NETD* augmente, le pourcentage d'images qui respect $\xi < 8°$ diminue, ce qui démontre l'influence du niveau de bruit temporel sur la précision de l'estimation. Nous remarquerons que la précision se dégrade assez lentement jusqu'à au moins une *NETD* simulée de 300 mK.

Remarque : le pourcentage d'images qui respecte $\xi < 8°$ n'évolue pas d'une manière monotone. Plusieurs aspects peuvent expliquer cela. Le premier est que l'algorithme « local », via la méthode RANSAC, fait intervenir un processus aléatoire. C'est-à-dire qu'on obtiendrait des résultats différents si on executait le code plusieurs fois sur une même image. Le second est que le bruit temporel est par nature un processus aléatoire, que nous avons simulé. Une analyse statistique permettrait sans doute d'obtenir une courbe plus lisse car elle diminurait l'impact de ces deux processus aléatoires.

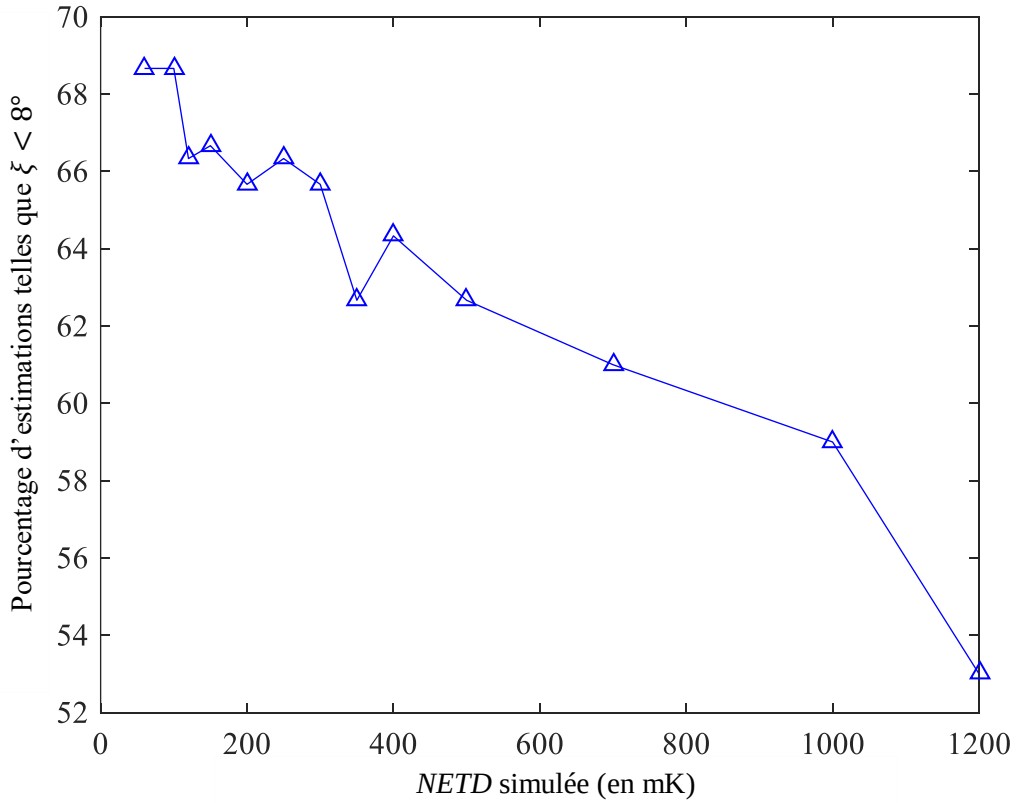


Figure 150. Pourcentage d'estimations de la pose dont les trois angles sont entachés d'une erreur inférieure à 8° en valeur absolue. Le format du détecteur est 640×480. La *NETD* réelle mesurée vaut environ 60 mK. Les images présentant une *NETD* supérieure ont été obtenues en simulant un bruit supplémentaire.

Nous avons ensuite testé l'influence d'une réduction du format du détecteur (taille et nombre de pixels) à champ de vue constant (53° selon la dimension horizontale). Là aussi nous nous appuyons sur l'estimation de ξ pour observer la dégradation de l'estimation de la pose. Nous avons donc recensé le pourcentage d'images au sein d'une vidéo de 300 images qui respecte la condition $\xi < 8°$. Seulement trois formats ont été testés : 640×480, 320×240 et 160×120 pixels. Le format 80×60 pixels ne permet pas de détecter et mettre en correspondance un nombre suffisant d'amers. Pour chacun des formats, nous avons simulé plusieurs *NETD* : 60, 100, 120, 150 et 200 mK. Les résultats, illustrés sur la Figure 151, montrent qu'une augmentation de la *NETD* se traduit par une diminution de la précision, à une exception près. En effet, on observe un « rebond » au niveau du point de mesure à 120 mK de *NETD* pour le format 320×240 pixels. Une analyse statistique permettrait sans doute d'obtenir des courbes monotones comme cela a été évoqué dans la précédente remarque.

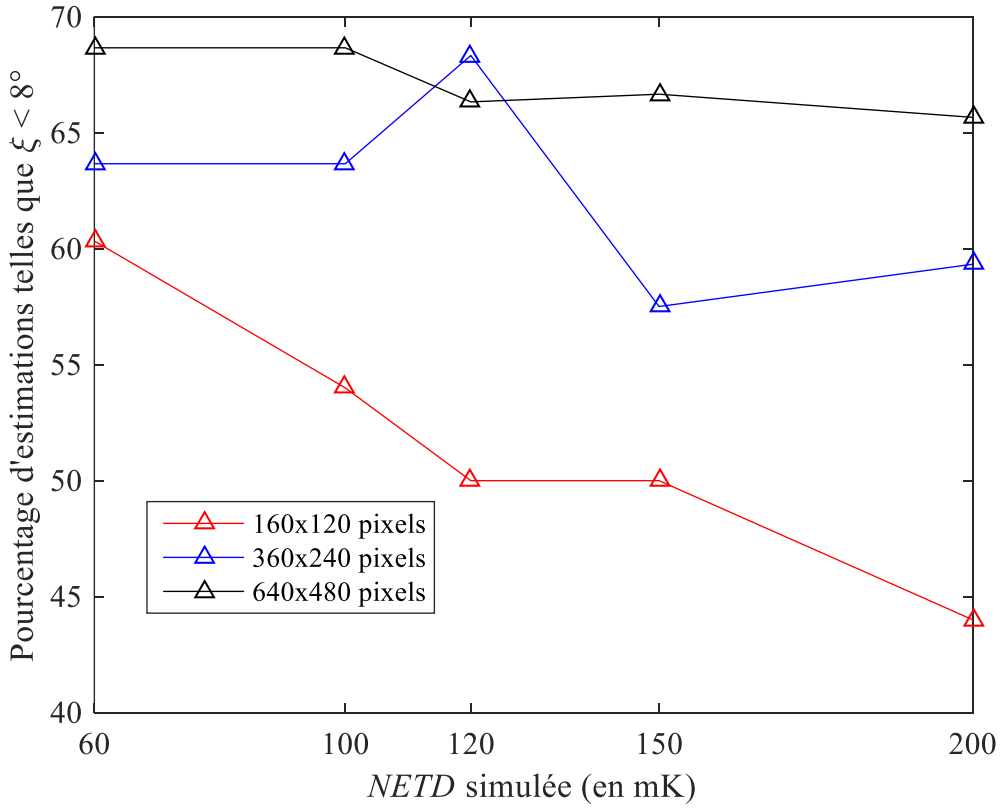


Figure 151. Pourcentage d'estimations de la pose dont les trois angles sont entachés d'une erreur inférieure à 8° en valeur absolue. Le détecteur a un format 160×120 pixels pour la courbe rouge et un format 320×240 pixels pour la courbe bleue. Le champ de vue est maintenu constant pour ces deux détecteurs (environ 53° selon l'axe horizontal). La *NETD* est artificiellement portée à 60 mK.

Concluons sur le fait que le champ de vue, assez resserré sur le visage dans nos expériences, doit très probablement être élargi si une telle application venait à se développer, afin de garantir que le visage soit toujours dans le champ de la caméra. Ainsi, si un algorithme « local » est utilisé, il semble très incertain de diminuer le format du détecteur pour gagner en coût car les performances d'estimations des angles sont très vites impactées. Par contre, et cela n'était pas acquis d'avance, si le format du détecteur est suffisamment grand, il semblerait possible de tolérer une *NETD* de l'ordre de 100 mK, même si ce constat doit être validée par une analyse statistique plus complète. Bien sur, toutes ces constatations ont du sens seulement dans le cadre de l'estimation de la pose et plus particulièrement de l'orientation du visage du conducteur.

6.5. L'algorithme « global »

Nous testons l'algorithme « global » de la même manière que l'algorithme « local ». Une seule vidéo de 300 images est utilisée dans laquelle le visage adopte des orientations typiques dans un contexte de conduite. Cette vidéo est acquise avec la caméra *Gobi* (format 640×480 pixels pour un champ de vue horizontal de 53° et avec une *NETD* d'environ 60 mK). Les autres formats et niveaux de *NETD* sont simulés à partir des images originales comme cela est détaillé dans les sections précédentes.

Comme nous l'avons fait pour l'algorithme « local », nous considérons que la caméra est étalonnée radiométriquement. Ainsi, nous utiliserons l'approche « globale » où l'on considère qu'il n'y pas d'offset entre les images de synthèse et les images réelles. Nous avons représenté tous les tests effectués avec l'approche « globale » sur la Figure 152. Chaque courbe représente un format en pixels. Et chaque point représente un niveau de *NETD*. Comme pour l'algorithme « local », pour une *NETD* et un format donné, nous estimons les 300 poses du visage au cours des 300 images de la vidéo et nous réalisons le pourcentage d'estimations dont l'erreur ξ' est inférieure à 8°. Précisons qu'avec l'algorithme « global », nous n'estimons pas l'angle *roll*. L'erreur utilisée ici est ξ' qui s'exprime comme suit :

$$\xi'(i) = \left|\max\left(e_{yaw}(i), e_{pitch}(i)\right)\right| \tag{6.13}$$

Le terme i permet de repérer l'image dans la vidéo ($i \in [1,300]$). Les termes $e_{yaw}(i)$ et $e_{pitch}(i)$ indiquent les erreurs commises sur les angles *yaw*, *pitch* et *roll* à la $i^{ième}$ image.

Il apparait très clairement sur Figure 152 que l'algorithme « global » conserve une précision importante du point de vue de l'estimation des angles *yaw* et *pitch* pour plusieurs formats et plusieurs niveaux de *NETD*. Sur cette même figure, on remarque que les formats les plus petits, c'est-à-dire les formats 160×120 et 80×60 pixels, avec une *NETD* de 250 mK, permettent d'estimer les angles *yaw* et *pitch* dans plus de 98% des images de la vidéo avec une précision meilleure que 8°. Cette remarque est évidement importante car les fabricants travaillent sur une réduction des coûts de ce type de capteurs. Ainsi, à la différence de l'algorithme « local », incapable de fonctionner avec un format 80×60 pixels, l'algorithme « global » semble beaucoup plus prometteur pour des applications grand publique.

Nous observons que les courbes sur la Figure 152 ne sont pas monotones. L'algorithme « global », différent de l'algorithme « local », ne fait pas intervenir un processus aléatoire dans son fonctionnement. Cependant le bruit temporel que nous avons simulé est un processus aléatoire. Nous expliquons la présence de ces rebonds observés par le caractère aléatoire du bruit temporel. Une analyse statistique permettrait de rendre ces courbes monotones. Nous avons mené cette analyse statistique pour le format 80×60 en executant dix fois l'estimation de la pose sur la vidéo de 300 images, et cela pour tous les niveaux de *NETD* simulés. Les points de mesure reportés sur la Figure 152 concernant ce format correspondent à une moyenne sur les dix tests. De plus, nous avons ajouté les barres d'erreurs dont la taille correspond au double de l'écart type. Cette fois ci, on constate une courbe monotone.

Il est très intéressant de constater qu'avec une *NETD* de 400 mK pour un format de 80×60 pixels, plus de 96 % des estimations sur les anlges *yaw* et *pitch* sont meilleures que 8°.

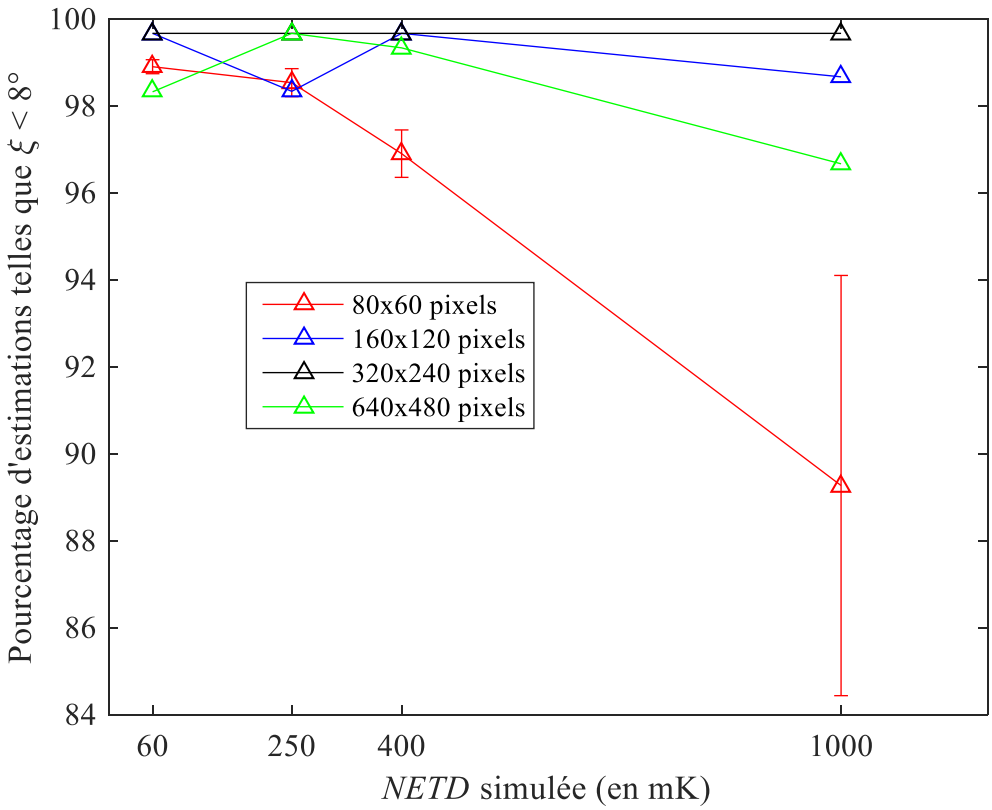


Figure 152. Pourcentage d'estimations des angles *yaw* et *pitch* entachés d'une erreur inférieure à 8° en valeur absolue. L'algorithme « global » a été testé pour différents formats et à différents niveau de *NETD*. Une seule vidéo de 300 images est utilisée dans laquelle le visage adopte des orientations typiques dans un contexte de conduite. Cette vidéo est acquise avec la caméra *Gobi* (format 640×480 pixels pour un champ de vue horizontal de 53° et avec une *NETD* d'environ 60 mK). Les autres formats et niveaux de *NETD* sont simulés à partir des images originales. Pour la courbe rouge avec des triangles qui représente le format 80×60 pixels, l'analyse de la vidéo a été effectuée dix foix. Chaque point de mesure est une moyenne sur dix executions du code. La taille des barres d'erreur correspond au double de l'écart type sur la mesure.

Les tests menés sur l'algorithme « global », jusqu'ici, ont été réalisés sur des vidéos acquises en laboratoire. Dans ces conditions, le visage est très bien démarqué du fond, qui rayonne beaucoup moins. Ce cas est favorable au bon fonctionnement de l'algorithme « global » car celui-ci cherche la zone de l'image réelle la plus proche des images de synthèse. Nous avons ajouté un fond artificiel dont le niveau correspond approximativement à celui de la peau. Le but, est de montrer que l'algorithme « global » est capable de fonctionner au-delà des conditions du laboratoire. Nous avons choisi de simuler une caméra au format 80×60 pixels.

La Figure 153 représente des images réelles et nos images simulées de capteurs à bas coût de spécifications proches en termes de *NETD* et de résolution spatiale. Cette figure démontre que les images que nous simulons sont relativement proches d'images réelles. De plus on s'aperçoit qu'en conditions réelles une journée chaude dans une automobile, la zone qui rayonne le plus de l'habitacle est la lunette arrière comme on peut le voir sur l'image a) de la Figure 153. Il nous semble assez improbable que tout

l'habitacle rayonne autant que la peau. Cependant nous simulons un cas où une très grande zone du fond rayonne autant que le visage (cf. d) de la Figure 153).

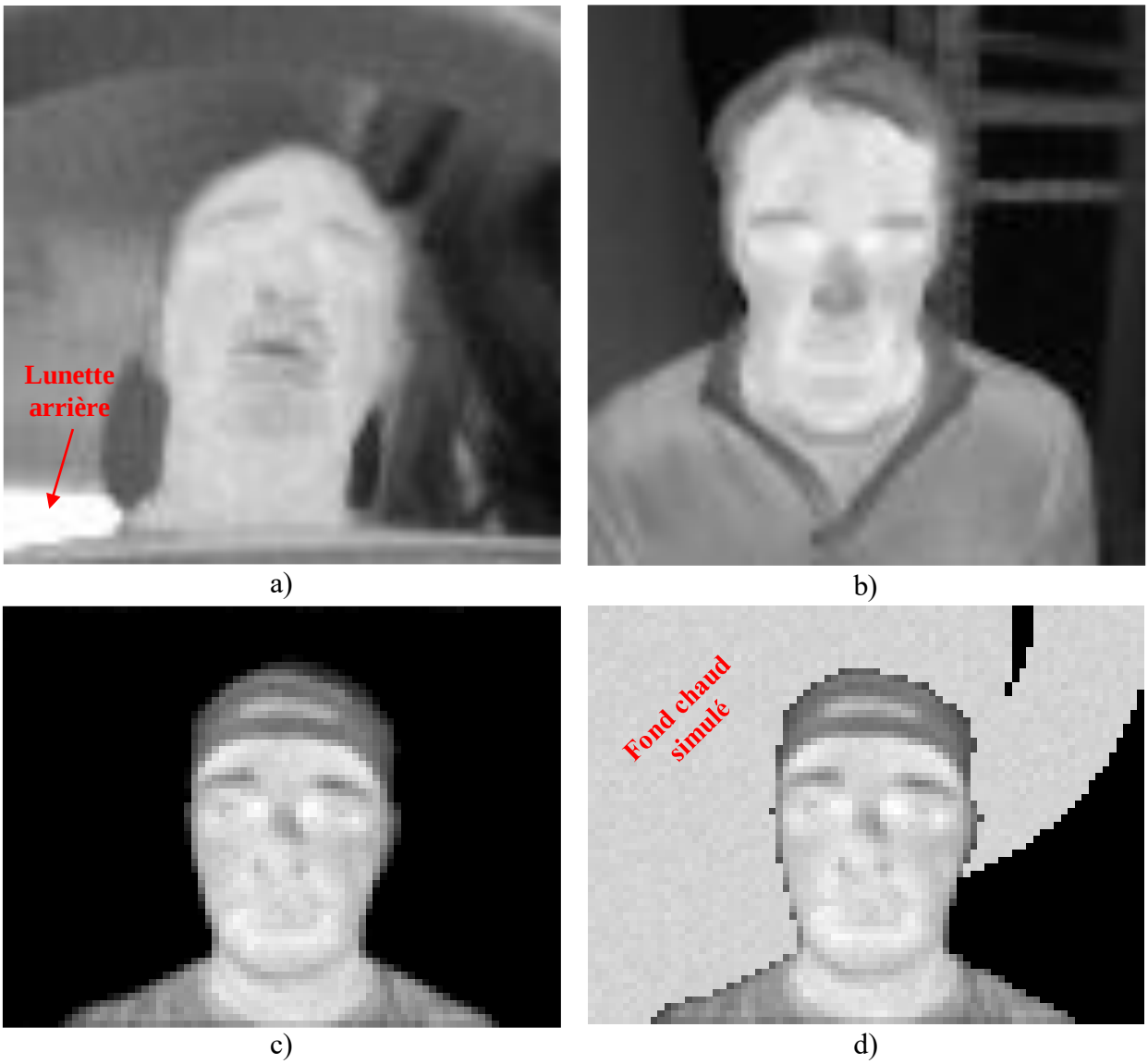


Figure 153. Comparaison visuelle d'images réelles et simulées d'un capteur à bas coût. a) Image réelle acquise une journée chaude avec une caméra *SmartIR80* de la société *Device-Alab* (format 80×80 pixels, champ de vue horizontal environ 50°, *NETD* de 119.5 mK) positionnée au niveau du compteur de vitesse, derrière le volant. b) Image réelle acquise avec la même caméra, en laboratoire. c) Image simulée (à partir d'une image réelle acquise avec la caméra *Gobi*) d'un format 80×60 pixels avec une NETD de 120 mK. d) image identique à l'image c) mais en ajoutant un fond chaud simulé.

Nous avons simulé plusieurs niveaux de *NETD* sur les images avec un fond chaud. A chaque niveau de *NETD* nous estimons la pose sur une vidéo d'approximativement 300 images et nous répétons dix fois cela pour obtenir une moyenne et un écart type sur l'erreur maximale ξ'. Les résultats sont illustrés sur la Figure 154. Ce test montre une seconde fois que, logiquement, la précision de l'estimation se dégrade lorsque la *NETD* augmente.

Nous observons également, en comparant les courbes avec et sans fond chaud, qu'en présence de ce dernier, la précision diminue. Cependant le cas testé nous semble être un cas difficile et la dégradation de la précision ne semble pas dramatique. En effet, pour les simulations où la *NETD* est inférieure à 400 mK, on observe une différence de l'ordre de deux pourcents entre les cas avec et sans fond chaud.

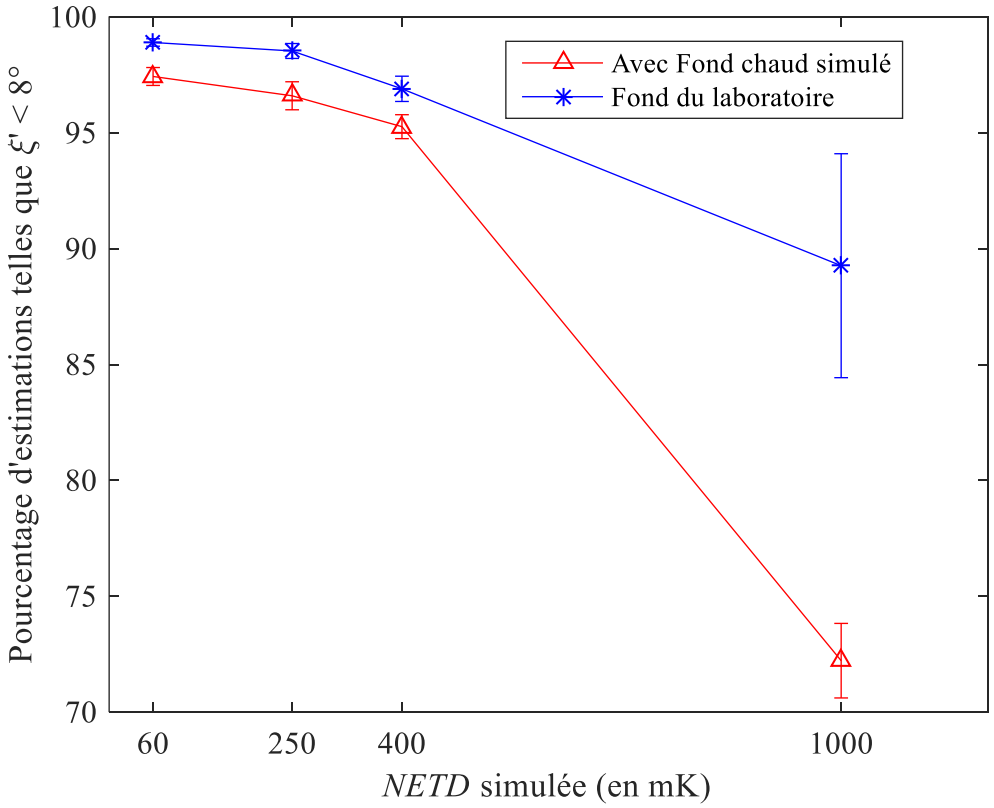


Figure 154. Pourcentage d'estimations des angles *yaw* et *pitch* dont l'erreur est inférieure à 8°. La courbe rouge avec des triangles correspond à une situation le fond de l'image rayonne approximativement autant que la peau. La courbe bleue avec des étoiles correspond à une situation où le fond rayonne beaucoup moins que le visage. Chaque point de mesure est une moyenne sur dix executions du code. La taille des barres d'erreur correspond au double de l'écart type sur la mesure.

La Figure 155 illustre certains résultats de l'algorithme « global » lorsqu'un fond chaud simulé est ajouté sur nos images simulées d'une caméra à bas coût. On s'apperçoit sur cette figure que l'algorithme « global » estime convenablement la position du visage, malgré la présence d'un fond chaud. Nous avons

Figure 155. Superpositions des images réelles et des images de synthèse (en rose) qui représentent le mieux l'orientation selon les angles *yaw* et *pitch*. Un format 80×60 et une *NETD* de 120 mK ont été simulés à partir de l'image réelle acquise par la caméra *Gobi*.

également évalué le pourcentage d'estimations pour lesquelles le maximum de l'erreur sur *yaw* et *pitch* noté ξ' est inférieur à 6° et nous obtenons 81.7% pour un format de 80×60 pixels et une *NETD* de 120 mK.

Le Tableau 14 donne l'erreur moyenne et l'écart type de l'erreur sur les angles *yaw* et *pitch* au cours d'une seule séquence vidéo. Le format simulé est 80×60 et la *NETD* simulée est 120 mK. Sur la première ligne du tableau, ce sont les résultats obtenus sans ajouter un fond chaud simulé. Sur le seconde du ligne du tableau, ce sont les résultats obtenus lorsqu'un fond chaud simulé est ajouté.

Tableau 14. Erreur moyenne et écart type de l'érreur commise sur les angles *yaw* et *pitch* avec un format 80×60 pixels pour un champ de vue horizntal de 53° avec une *NETD* de 120 mK.

Résultats des estimations sur les rotations ($EM \pm S$)				
Fond de l'image	**Format (en pixels×pixels)**	**NETD (en mK)**	**Erreur sur *pitch***	**Erreur sur *yaw***
Laboratoire	80×60	120	0.2° ± 3,2°	-0.7° ± 3.9°
Fond chaud simulé	80×60	120	0.3° ± 3,8°	-1.7° ± 3.7°

La comparaison des erreurs moyennes et des écarts types sur les erreurs des angles *yaw* et *pitch* avec et sans fond chaud simulé illustre, sur une vidéo, que l'algorithme « global » est impacté assez faiblement par un fond chaud, qui est le cas critique pour l'algorithme « global ».

6.6. Conclusion

Si l'on considère uniquement des applications d'estimation de la pose (ou d'une partie de la pose comme pour l'algorithme « global »), les résultats présentés dans ce chapitre montrent que la *NETD* doit être spécifiée en accord avec la précision requise sur l'estimation des angles. Elle doit également tenir compte de l'algorithme qui va être implémenté (« global » ou « local »). Dans tous les cas, une *NETD* de 100 mK nous semble suffisante aux vus des méthodes implémentées et de l'application visée. Evidemment pour d'autres applications, d'autres tests doivent être menés.

D'après les expériences menées dans ce chapitre, l'algorithme « global » semble mieux adapté que l'algorithme « local » aux basses résolutions spatiales, ce qui est en accord avec la littérature déjà introduite au chapitre 1 [131,135]. Ce que nous n'avions pas prédit, c'est que l'algorithme « global » semble robuste à des niveaux de *NETD* importants. Ces premiers résultats amènent à penser que l'algorithme « global » fonctionnerait avec des caméras considérées comme bas coût. Bien sûr une analyse statistique doit être menée pour confirmer cela, ainsi qu'une étude avec une caméra bas coût réelle. Notons que dans cette expérience, nous ne tenons pas compte des abérations optiques qui interviendrons lorsqu'une caméra bas cout réelle sera utilisée.

Le bénéfice d'un fonctionnement à basse résolution spatiale est double : le prix du capteur diminue, le temps de calcul est diminué également.

Précisons une fois de plus (car nous l'avons déjà précisé au chapitre 4) que l'algorithme « global » estime les angles *yaw* et *pitch* ainsi que les translations dans le plan. Il manque donc l'angle *roll* ainsi qu'une

translation (distance entre la caméra et le visage) pour estimer la pose complète du visage. Nous avons déjà mentionné des articles permettant d'estimer l'angle *roll* [68]. L'estimation de la distance entre la caméra et le visage peut être réalisée en ajoutant des images à la base d'images de synthèse ce qui augmenterait le temps de calcul. Pour contrer cela, l'utilisation de l'accumulation temporelle des estimations peut être exploitée (grâce à un filtre de Kalman par exemple).

Conclusion et Perspectives

Conclusion

Les travaux de recherche présentés dans ce manuscrit ont pour objectif d'étudier l'apport de l'imagerie thermique pour l'estimation de pose du visage d'un conducteur. Le sujet étant nouveau pour l'entreprise, chez qui j'ai fait ma thèse, j'ai équipé l'un des laboratoires de Renault avec le matériel nécessaire. Grâce à une collaboration avec l'*ONERA*, j'ai pu reproduire et adapter à mon besoin, un banc d'acquisition d'images très simple. Grâce à ce banc, j'ai étudié et évalué différentes méthodes de correction du *BSF* : correction basée sur un obturateur mécanique, correction basée sur l'étalonnage en chambre climatique, correction basée sur des méthodes de traitement d'images.

Suite à cette évaluation, j'ai choisi de corriger le *BSF* grâce aux corrections les plus efficaces, c'est-à-dire l'étalonnage en chambre climatique et l'obturateur mécanique, tout en sachant qu'elles ne sont pas nécessairement compatibles avec les contraintes automobiles. Néanmoins, les progrès des algorithmes par traitement d'images semblent encourageant pour la suite et permettrons d'envisager des alternatives à l'obturateur mécanique et à l'étalonnage en chambre climatique.

Grâce au travail que j'ai mené pour pouvoir synchroniser les images thermiques avec une vérité terrain pour l'estimation de la pose du visage, j'ai pu créer une base d'images thermiques labellisées en orientation (angles *yaw, pitch* et *roll*). Cette base d'images est composée de vingt vidéos acquises pour des personnes de couleur de peau, d'âge et de sexe différents. J'ai veillé à acquérir ces vidéos dans un contexte automobile grâce à un modèle simplifié d'un *Renault Scenic* et à un scénario toujours identique.

J'ai ensuite focalisé le sujet de recherche sur la précision de l'estimation de la pose du visage par caméra thermique car j'ai jugé que cette brique était essentielle pour envisager d'autres applications relatives à la surveillance du conducteur. Pour cela, je me suis orienté vers des approches de l'état de l'art utilisant des maillages. J'ai pu tester deux maillages 2D du visage, le premier a été obtenu grâce à un scan 3D du visage, le second est un ellipsoïde décrivant approximativement la géométrie 3D du visage. À partir de ces deux maillages, j'ai pu évaluer l'influence de la précision des modèles dans la mesure de pose du visage.

Grâce à la collaboration avec le Laboratoire Hubert Curien, j'ai appliqué les méthodes d'estimation de la pose, développées initialement pour l'imagerie visible. Deux algorithmes qui utilisent un maillage 3D du visage ont été implémentés. Ces deux algorithmes sont opposés par leur mode de fonctionnement. Le premier est caractérisé de « local » car il nécessite la détection et la description de points d'intérêt. Le second est dit « global » car tous les pixels du visage sont utilisés pour l'estimation de la pose.

Ces deux méthodes de traitement d'images ont été évaluées et comparées entre elles. J'ai montré que l'étalonnage radiométrique permettait de réduire les temps de calcul des deux méthodes. J'ai constaté qu'une précision plus importante sur les angles *yaw* et *pitch* est plus facilement atteignable avec la méthode « globale » qu'avec la méthode « locale » avec une caméra relativement couteuse.

Un brevet sur une procédure d'estimation de la pose du visage par imagerie thermique basée sur la méthode « locale » a été déposé. Une conférence internationale avec acte de soutenance orale m'a également permis de présenter les premiers résultats obtenues avec les méthodes « globale » et « locale ».

Enfin, j'ai mené une étude sur l'influence de la réduction de la résolution spatiale et de l'augmentation de la *NETD* sur la précision d'estimation de l'orientation. J'ai ainsi pu mettre en avant les avantages de l'algorithme « global », qui est plus robuste que l'algorithme « local », aux basses résolutions spatiales et à des niveaux de *NETD* importants.

Perspectives

Une partie de la faisabilité de l'estimation de la pose du visage par caméra thermique a été démontrée dans ces travaux. Néanmoins, la phase de plaquage de texture, que j'ai réalisée manuellement, doit être automatisée. La base de vidéos thermiques labellisées que j'ai créée peut être utilisée pour valider l'ensemble de la chaîne algorithmique, c'est-à-dire plaquage de texture automatisé et estimation de la pose. Une fois le système validé en laboratoire, pour une résolution thermique et un format donné, une validation réelle couvrant un panel de cas d'usages bien choisi doit être effectuée.

La surveillance du conducteur ne se restreint pas seulement à l'estimation de la pose. J'ai présenté dans le chapitre 1, différentes méthodes basées sur la mesure de la température de zones particulières du visage pour évaluer le niveau de stress ou de fatigue du conducteur. Les méthodes présentées dans ces travaux de thèse permettront de suivre des zones du visage intéressantes à contrôler en température. Ainsi, des travaux sur la construction d'un indicateur de stress ou de fatigue basé sur le suivi de la température pourront être envisagés.

Il sera intéressant d'étudier la combinaison de plusieurs modalités telles que l'imagerie thermique et l'imagerie visible/*NIR*. Les avantages des deux modalités pourront être cumulés :

- la robustesse face aux variations d'illumination et la possibilité de suivre la température pour la caméra thermique ;
- l'information plus riche en hautes fréquences spatiales et la possibilité d'évaluer l'état des yeux (ouverts ou fermés) pour la caméra visible.

Certains travaux permettant d'aligner des images d'un visage acquises dans ces deux modalités (thermique et visible/NIR) ont déjà été publiés [168,169168]. La configuration, caméra visible/*NIR* plus caméra thermique, ne résout cependant pas le problème de l'illumination du visage du conducteur par LEDs *NIR* si le système global nécessite qu'elles soient toujours allumées. Une diminution du temps d'illumination

du visage pourrait alors faire l'objet d'une réflexion en se basant sur l'hypothèse que dans la majorité des scénarios, la caméra thermique seule suffit.

A plus long terme, les progrès des détecteurs thermiques permettront d'imager le réseau vasculaire présent sous la peau, ce qui créera naturellement de la texture facilitant sans aucun doute les algorithmes d'estimation de la pose et de suivi ainsi que l'extraction de paramètres physiologiques. Des approches telles que les travaux du *MIT (Massachusetts Institute of Technology) Lincoln Laboratory* proposent d'utiliser un cristal liquide pour convertir le rayonnement thermique en rayonnement visible pour obtenir un imageur possédant un grand format ainsi qu'une bonne résolution thermique [170]. D'autres travaux tels que ceux du *CEA Leti* sur les structures vibrantes *NEMS* (*Nano Electro Mechanical Systems*) permettent de réduire le bruit temporel [171].

Annexes

Annexe A : bases radiométriques

La loi de Planck exprime la luminance spectrique (en $W.m^{-2}.sr^{-1}.m^{-1}$) d'un corps noir en fonction de sa température T_{CN} :

$$L_\lambda = \frac{2hc^2}{\lambda^5} \frac{1}{exp\left(\frac{hc}{\lambda k_B T_{CN}}\right) - 1} \quad \text{(A.1)}$$

c : vitesse de la lumière dans le milieu considéré ($c = c_0/n$) avec c_0 vitesse de la lumière dans le vide et n indice du matériau. $c_0 = 299\,792\,458\ m/s$.
k_B : constante de Boltzman, $k = 1{,}38.10^{-23}\ J/K$.
h : constante de Planck, $h = 6{,}63.10^{-34}\ J.s$.
λ : Longueur d'onde du rayonnement (en m).

La bande spectrale de détection des bolomètres est 8-14 µm. Nous définissons la notation de l'intégrale $L_{T_{CN}}^{8-14}$ entre 8 et 14 µm de la luminance spectrale à la température T_{CN} comme suit :

$$L_{T_{CN}}^{8-14} = \int_{8\mu m}^{14\mu m} L_\lambda d\lambda \quad \text{(A.2)}$$

Dans ce projet, nous travaillons toujours avec des détecteurs sensibles dans la bande spectrale 8-14 µm, nous nous permettons donc d'alléger la notation $L_{T_{CN}}^{8-14}$ par $L_{T_{CN}}$.

Le flux $F_{scène}$ en watt reçu par un détecteur et rayonné par la scène :

$$F_{scène} = \varepsilon \tau_{atm} \tau_{optique} L_{scène} G \quad \text{(A.3)}$$

La transmission de l'optique, intégrée sur la bande spectrale du détecteur, est notée $\tau_{optique}$. De même, la transmission de l'atmosphère, intégrée sur la bande spectrale du détecteur, est notée τ_{atm}. L'étendue géométrique est notée G. Pour l'évaluer nous devons considérer le champ de réception du détecteur. Il est souvent noté $iFOV$ (*Instantaneous Field of View*). Lorsque que le champ de réception du détecteur $iFOV$ est inférieur aux dimensions de la source on dit que le détecteur est un capteur d'image (dans le cas inverse, on parle de capteur de flux). On peut calculer l'étendue géométrique comme suit (cf. Figure 156) :

$$G = \pi S_{pixel}\, sin^2(\alpha') \quad \text{(A.4)}$$

La surface du pixel carré est notée S_{pixel}. Le demi angle d'ouverture de l'optique est notée α'.

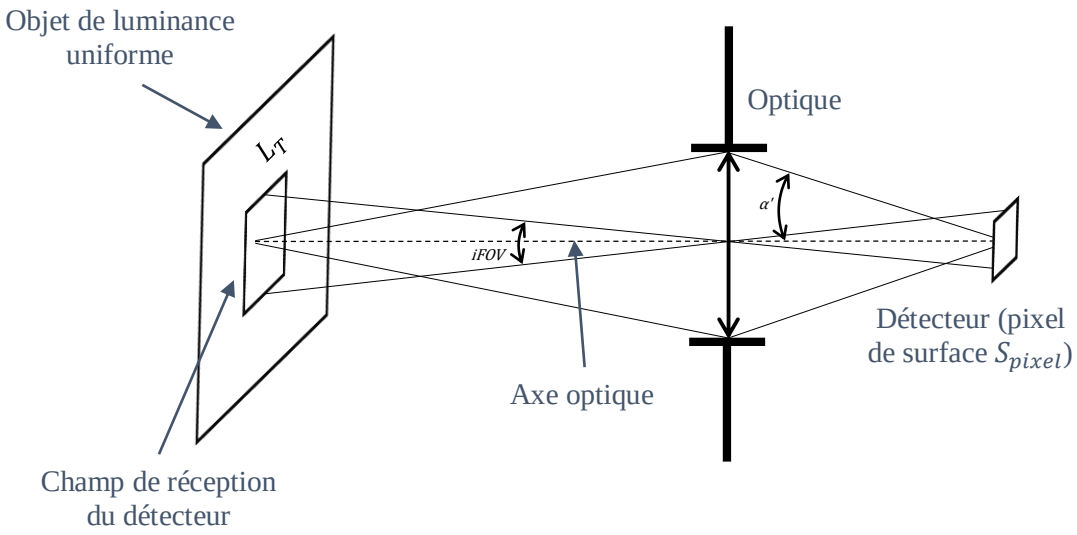


Figure 156. Schéma pour le calcul du flux reçu par un pixel sur l'axe optique (le rayonnement se propage de la gauche vers la droite).

Remarque sur le calcul de l'ouverture numérique :

Parfois, le nombre d'ouverture image N' est renseigné par le fabricant de caméra, mais pas l'ouverture numérique image $\sin(\alpha')$. *Dans le cas d'un système aplanétique, ces deux grandeurs sont liées par la relation suivante :*

$$N' = \frac{1}{2\sin(\alpha')}$$

L'éclairement des pixels est généralement atténué du centre vers les bords de la matrice à cause du vignettage. On ajoute généralement une fonction $\xi(\theta)$. Le gain d'une *NUC* « deux points » corrige cette atténuation d'éclairement.

$$F_{scène} = \tau_{atm}\tau_{optique}\pi S_{pixel}L_{scène}\, sin^2(\alpha')\,\xi(\theta) \quad \text{(A.5)}$$

Annexe B : calcul théorique de la *NETD* à partir de la mesure expérimentale du bruit

La définition générale de la *NETD* est :

$$NETD = \frac{\tilde{v}}{\Delta V}\Delta T$$

En remarquant que :

$$NETD = \tilde{v}\frac{\Delta T}{\Delta F}\frac{\Delta F}{\Delta V}, \tag{B.1}$$

et, en utilisant la forme différentielle de l'équation ci-dessus, on peut faire apparaître la *responsivité* en V/W :

$$NETD = \frac{\tilde{v}}{\Re}\frac{1}{\frac{\partial F}{\partial T}} \tag{B.2}$$

Le terme $\frac{\partial F}{\partial T}$ s'exprime analytiquement :

$$\frac{\partial F}{\partial T} = \frac{\pi S_{pixel}}{4N'^2}Q_T \tag{B.3}$$

Le terme Q_T correspond à :

$$Q_T = \frac{\partial L_{scène}}{\partial T} = \int_{8\mu m}^{14\mu m} \frac{\partial L_\lambda}{\partial T}\, d\lambda \ , \tag{B.4}$$

avec,

$$\frac{\partial L_\lambda}{\partial T} = L_\lambda \frac{hc}{\lambda T^2 k_B}\frac{1}{1 - exp\left(-\frac{hc}{\lambda k_B T_{CN}}\right)}. \tag{B.5}$$

En combinant les équations (B.2) et (B.3), on peut exprimer la *NETD* à partir du bruit en V, de la *responsivité* en V/W et des paramètres opto-mécaniques.

$$NETD = \frac{4N'^2}{S_{pixel}}\frac{1}{Q_T}\frac{\tilde{v}}{\Re_V} \tag{B.6}$$

Annexe C : prise en compte d'un offset global entre les images de synthèse et l'image réelle

On considère que les pixels Y_{ij} du visage de l'image réelle sont égaux à ceux des pixels g_{ij} du visage de l'image de synthèse à un bruit près ε_{ij} et à un offset global près β :

$$Y_{ij} = g_{ij} + \varepsilon_{ij} + \beta \tag{C.1}$$

Nous cherchons le décalage entre la position du visage dans l'image réelle et dans l'image de synthèse et l'offset β par une approche « problèmes inverses ». La fonction de coût à minimiser est :

$$\mathcal{P}(u,v) = \sum_{ij} w_{uvij}\left(Y_{ij} - g_{uvij} - \beta\right)^2 \tag{C.2}$$

En développant le carré on obtient :

$$\begin{aligned}\mathcal{P}(u,v) = &\sum_{ij} w_{uvij} Y_{ij}^2 + \sum_{ij} w_{uvij} g_{ij}^2 + \sum_{ij} w_{uvij}\beta^2 \\ &-2\sum_{ij} w_{uvij} Y_{ij} g_{uvij} - 2\sum_{ij} w_{uvij} Y_{ij}\beta - 2\sum_{ij} w_{uvij} g_{uvij}\beta\end{aligned} \tag{C.3}$$

L'offset β^+ qui minimise la fonction de coût est tel que :

$$\frac{\partial \mathcal{P}}{\partial \beta}\Big)_{\beta^+} = 0 \tag{C.4}$$

Exprimons $\frac{\partial \mathcal{P}}{\partial \beta}$ à l'aide de l'équation (C.3) :

$$\frac{\partial \mathcal{P}}{\partial \beta} = 2\sum_{ij} w_{uvij}\beta - 2\sum_{ij} w_{uvij} Y_{ij} - 2\sum_{ij} w_{uvij} g_{uvij} \tag{C.5}$$

D'après les équations (C.4) et (C.5)L'expression de β^+ est donc :

$$\beta^+ = \frac{\sum wY(u) + \sum wg}{\sum w} \tag{C.6}$$

Dans l'équation (C.6), des changements de notations ont été effectués afin d'alléger les écritures. Les produits de corrélations apparaissent grâce à une dépendance à la variable (u) de l'une des fonctions :

$$\begin{aligned}\sum wg &= \sum w_{uvij} g_{uvij} \\ \sum w &= \sum w_{uvij} \\ \sum wY(u) &= \sum w_{uvij} Y_{ij} \qquad \rightarrow \text{Produit de corrélation}\end{aligned}$$

L'estimation de β^+ nécessite donc le calcul d'un produit de correlation.

Nous effectuerons également les changements de notations suivant :

$\sum wY^2(u) = \sum w_{uvij} Y_{ij}^2$ → Produit de corrélation
$\sum wg^2 = \sum w_{uvij} g_{uvij}^2$
$\sum wgY(u) = \sum w_{uvij} Y_{ij} g_{uvij}$ → Produit de corrélation

Insérons l'expression maintenant l'expression de β^+ dans l'équation (C.3) :

$$\begin{aligned}\mathcal{P}(u,v)\big)_{\beta=\beta^+} = \sum wY^2(u) + \sum wg^2 + \left(\frac{\sum wY(u) + \sum wg}{\sum w}\right)^2 \sum w \\ -2\sum wgY(u) - 2\left(\frac{\sum wY(u) + \sum wg}{\sum w}\right).\left(\sum wY(u) - \sum wg\right)\end{aligned} \quad \text{(C.7)}$$

Après quelques manipulations on obtient l'expression suivante :

$$\begin{aligned}\mathcal{P}(u,v)\big)_{\beta=\beta^+} = \sum wY^2(u) + \sum wg^2 - 2\sum wgY(u) \\ -\frac{(\sum wY(u) - \sum wg)^2}{\sum w}\end{aligned} \quad \text{(C.8)}$$

L'équation (C.8) contient trois produits de corrélations. Sachant que trois *FFTs* sont nécessaires pour calculer un produit de corrélation, cette équation requiert le calcul de neuf *FFTs*.

Annexe D : *parabolic estimator*

La fonction à deux variables $\mathcal{P}(yaw, pitch)$ échantillonné avec un pas de 5° est minimum pour le couple $yaw_{min},\ pitch_{min}$. Nous cherchons à affiner l'estimation$yaw_{min},\ pitch_{min}$.

Nous admettons que les variables de la fonction $\mathcal{P}(yaw, pitch)$ sont séparables grâce à deux fonctions $\mathcal{P}_{yaw}$ et $\mathcal{P}_{pitch}$ telle que :

$$\mathcal{P}(yaw, pitch) = \mathcal{P}_{yaw}(yaw)\mathcal{P}_{pitch}(pitch) \tag{D.1}$$

Nous réalisons une régression polynomiale d'ordre 2 de la fonction $\mathcal{P}_{yaw}(yaw)$ autour du voisinage local de yaw_{min}. On réalise la régression grâce à cinq valeurs dont la centrale est yaw_{min}. On extrait ainsi trois coefficients. La fonction $\mathcal{P}_{yaw}$ peut être approximée ainsi :

$$\mathcal{P}_{yaw}(yaw) = a.(yaw)^2 + b.yaw + c \tag{D.2}$$

Puis, en étudiant la pente de la fonction $\mathcal{P}_{yaw}(yaw)$, on obtient une estimation $\mathcal{P}_{yaw}(\widehat{yaw}_{min})$ du minimum qui n'est pas limité par le pas d'échantillonnage (on cherche yaw_{min} est tel que $\frac{\partial \mathcal{P}_{yaw}}{\partial yaw}\Big)_{yaw_{min}} = 0$) :

$$\widehat{yaw}_{min} = -\frac{b}{2a} \tag{D.3}$$

Remarque : Cette technique est très adaptée pour des pics symétriques.

En effectuant la même opération avec la fonction $\mathcal{P}_{pitch}(pitch)$ on obtient une estimation du minimum de la fonction $\mathcal{P}_{pitch}(\widehat{pitch}_{min})$ qui n'est pas limité par le pas d'échantillonnage. Enfin, on considère que la fonction $\mathcal{P}$ atteint un minimum pour les variables $(\widehat{yaw}_{min}, \widehat{pitch}_{min})$.

Nous ferons référence à cette technique classique grâce au terme *parabolic estimation* que l'on trouve dans la référence [133].

Annexe E : quantité de bruit en *Adu* ajouté pour simulé des niveaux de *NETD* élevés

Cette annexe montre le bruit ajouté à l'image pour simuler *NETD* comme cela est expliqué au Chapitre 6 à la section 6.3. La première colonne est le format en pixels que nous simulons grâce à un filtrage d'un pixel équivalent comme cela est expliqué dans la section 6.2. La deuxième colonne correspond à la *NETD* que nous souhaitons obtenir par simulation. La troisième colonne correspond à l'équivalent en *Adu* du bruit temporel. La seconde et la troisième colonne sont liées par la *responsivité* $\mathfrak{R}_{Adu/K}$ en *Adu*/K comme suit :

$$\sigma_{ciblée} = NETD_{ciblée} \times \langle \mathfrak{R}_{Adu/K} \rangle \quad \text{(E.1)}$$

La quatrième colonne correspond à la *NETD*, en mK, évaluée après avoir appliqué le filtrage du pixel équivalent sur des images du corps noir. La cinquième colonne correspond au bruit en *Adu* après l'étape de filtrage du pixel équivalent. La sixième colonne correspond au bruit que nous ajoutons et qui est calculé grâce à l'équation suivante :

$$\sigma^2_{ajouté} = \sigma^2_{ciblé} - \sigma^2_{lissé} \quad \text{(E.2)}$$

Enfin, la dernière colonne correspond à une évaluation de la *NETD* après avoir ajouté le bruit $\sigma_{ajouté}$ à des images du corps noir ayant subies le filtrage du pixel équivalent.

Tableau 15. Récapitulatif des configurations testées.

Format (en pixels× pixels)	*NETD* ciblée (en mK)	$\sigma_{ciblée}$ (en Adu)	*NETD* après lissage (en mK)	$\sigma_{lissé}$ (en *Adu*)	$\sigma_{ajouté}$ (en *Adu*)	*NETD* après ajout du bruit simulé (en *Adu*)*
Les images originales sont acquises avec un capteur au 640×480 pixels. La médiane de la *NETD* mesurée de la caméra vaut 58 mK						
80×60	60	10.3	22.2	3.9	9.5	59.1
80×60	100	17.2	22.2	3.9	16.8	99.6
80×60	120	20.6	22.2	3.9	20.2	118.4
80×60	150	25.8	22.2	3.9	25.5	149
80×60	200	34.4	22.2	3.9	34.2	198.3
80×60	250	43	22.2	3.9	42.8	247.4
80×60	400	68.8	22.2	3.9	68.7	397.8
80×60	1000	172.1	22.2	3.9	172.1	996.6
160×120	60	10.3	26.3	4.6	9.2	59.2
160×120	100	17.2	26.3	4.6	16.6	99.2
160×120	120	20.6	26.3	4.6	20.1	118.8

160×120	150	25.8	26.3	4.6	25.4	149
160×120	200	34.4	26.3	4.6	34.1	198.4
160×120	250	43	26.3	4.6	42.8	258.6
160×120	400	68.3	26.3	4.6	68.1	393.8
160×120	1000	172.1	26.3	4.6	172	991.5

320×240	60	10.3	36.3	6.3	8.1	53.7
320×240	100	17.2	36.3	6.3	16.0	96
320×240	120	20.6	36.3	6.3	19.6	115.9
320×240	150	25.8	36.3	6.3	25.0	146.6
320×240	200	34.4	36.3	6.3	33.8	197.2
320×240	250	43.02	36.3	6.3	42.6	248.3
320×240	400	68.8	36.3	6.3	68.5	396.6
320×240	1000	172.1	36.3	6.3	172	992.8

640×480	100	17.2	58**	10**	14	99.5
640×480	120	20.6	58**	10**	18	118.9
640×480	150	25.8	58**	10**	23.8	149
640×480	200	34.4	58**	10**	32.9	198.4
640×480	250	43	58**	10**	41.8	248
640×480	300	51.6	58**	10**	50.6	297.6
640×480	350	60.2	58**	10**	59.4	347.5
640×480	400	68.8	58**	10**	68.1	397.1
640×480	450	77.4	58**	10**	76.8	446.8
640×480	500	86	58**	10**	85.5	496.5
640×480	700	120.5	58**	10**	120.1	694.9
640×480	1000	172.1	58**	10**	170.8	987.1
640×480	1200	206.5	58**	10**	206.3	1191.1

* Nous souhaitons que la *NETD* après ajout du bruit simulé (dernière colonne du tableau ci-dessus) soit proche de la *NETD* ciblée deuxième colonne du tableau ci-dessus).

** L'image n'est pas lissé car il s'agit dur format d'origine du capteur. Le bruit temporel de 10 *Adu* est celui de la caméra *Gobi*.

Bibliographie

1. Hoch, S., Schweigert, M., Althoff, F., & Rigoll, G. (2007, June). The BMW SURF project: A contribution to the research on cognitive vehicles. In *Intelligent Vehicles Symposium, 2007 IEEE* (pp. 692-697). IEEE.
2. Tawari, A., Sivaraman, S., Trivedi, M. M., Shannon, T., & Tippelhofer, M. (2014, June). Looking-in and looking-out vision for urban intelligent assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking. In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE* (pp. 115-120). IEEE.
3. Tawari, A., & Trivedi, M. M. (2014, June). Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos. In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE* (pp. 344-349). IEEE.
4. Doshi, A., & Trivedi, M. M. (2009). On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. *IEEE Transactions on Intelligent Transportation Systems*, *10*(3), 453-462.
5. Lam, C. P., Yang, A. Y., Driggs-Campbell, K., Bajcsy, R., & Sastry, S. S. (2015, September). Improving human-in-the-loop decision making in multi-mode driver assistance systems using hidden mode stochastic hybrid systems. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on* (pp. 5776-5783). IEEE.
6. Čolić, A., Marques, O., & Furht, B. (2014). *Driver Drowsiness Detection: Systems and Solutions*, (Introduction, page 2). Springer International Publishing.
7. World Health Organization. Violence, Injury Prevention, & World Health Organization. (2013). *Global status report on road safety 2013: supporting a decade of action*, (Résumé). World Health Organization.
8. Institut National du Sommeil et de la Vigilance (INSV) & Association Professionnelle Autoroutes et Ouvrages Routiers (ASFA). (2013). *Somnolence au volant – Le livre blanc* – Synthèse.
9. Knipling, R. R., & Wang, J. S. (1994). *Crashes and fatalities related to driver drowsiness/fatigue*. Washington, DC: National Highway Traffic Safety Administration.
10. http://media.daimler.com/marsMediaSite/en/instance/ko/ATTENTION-ASSIST-Drowsiness-detection-system-warns-drivers-t.xhtml?oid=9361586
11. http://www.volkswagen.co.uk/technology/passive-safety/driver-alert-system
12. Čolić, A., Marques, O., & Furht, B. (2014, August). Design and implementation of a driver drowsiness detection system: A practical approach. In *Signal Processing and Multimedia Applications (SIGMAP), 2014 International Conference on* (pp. 241-247). IEEE.
13. Sun, Y., Yu, X., Berilla, J., Liu, Z., & Wu, G. (2011). An in-vehicle physiological signal monitoring system for driver fatigue detection. In *3rd International Conference on Road Safety and Simulation*.

14. Solaz, J., Laparra-Hernández, J., Bande, D., Rodríguez, N., Veleff, S., Gerpe, J., & Medina, E. (2016). Drowsiness detection based on the analysis of breathing rate obtained from real-time image recognition. *Transportation Research Procedia*, *14*, 3867-3876.
15. Hartley, L. R., Arnold, P. K., Smythe, G., & Hansen, J. (1994). Indicators of fatigue in truck drivers. *Applied ergonomics*, *25*(3), 143-156.
16. Vicente, J., Laguna, P., Bartra, A., & Bailón, R. (2016). Drowsiness detection using heart rate variability. *Medical & biological engineering & computing*, *54*(6), 927-937.
17. http://www.bmw.com/com/en/insights/technology/technology_guide/articles/head_up_display.html
18. Watanabe, H., Yoo, H., Tsimhoni, O., & Green, P. (1999, November). The effect of HUD warning location on driver responses. In *Sixth World congress on Intelligent Transport Systems, Toronto, Canada*.
19. Doshi, A., Cheng, S. Y., & Trivedi, M. M. (2009). A novel active heads-up display for driver assistance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(1), 85-93.
20. http://www.greencarcongress.com/2017/02/20170227-transeev.html#comments
21. Rahman, H., Begum, S., & Ahmed, M. U. (2015, November). Driver Monitoring in the Context of Autonomous Vehicle. In *SCAI* (pp. 108-117).
22. http://smarteye.se/wp-content/uploads/2014/12/Product-Sheet-AntiSleep.pdf
23. http://www.innov-plus.com/fr/toucango/
24. Li, S. Z., Chu, R., Liao, S., & Zhang, L. (2007). Illumination invariant face recognition using near-infrared images. *IEEE Transactions on pattern analysis and machine intelligence*, *29*(4).
25. Heinzmann, J., & Zelinsky, A. (1998, April). 3-D facial pose and gaze point estimation using a robust real-time tracking paradigm. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on* (pp. 142-147). IEEE.
26. Lee, S. E., Olsen, E. C., & Wierwille, W. W. (2004). *A comprehensive examination of naturalistic lane-changes* (No. HS-809 702,).
27. Lépine, T., & Meyzonnette, J. L. (2001). Bases de radiométrie optique. *Cépaduès Editions*.
28. Boyd, R. W. (1983). Radiometry and the detection of optical radiation.
29. Soerensen, D. D., Clausen, S., Mercer, J. B., & Pedersen, L. J. (2014). Determining the emissivity of pig skin for accurate infrared thermography. *Computers and Electronics in Agriculture*, *109*, 52-58.
30. Feldstein, I., Guentner, A., & Bengler, K. (2015). Infrared-Based In-Vehicle Head-Tracking. *Procedia Manufacturing*, *3*, 829-836.
31. Chang, H., Harishwaran, H., Yi, M., Koschan, A., Abidi, B., & Abidi, M. (2006, June). An indoor and outdoor, multimodal, multispectral and multi-illuminant database for face recognition. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on* (pp. 54-54). IEEE.
32. Trivedi, M. M., Cheng, S. Y., Childers, E. M., & Krotosky, S. J. (2004). Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation. *IEEE transactions on vehicular technology*, *53*(6), 1698-1712.
33. http://www.yole.fr/IR_Detector_Technologies.aspx#.WMQiF6ozU5h

34. Bodo, F. O. R. G., Herrmann, F., Leneke, W., Schieferdecker, J., Simon, M., Storck, K., & Schulze, M. (2013). *U.S. Patent No. 8,592,765*. Washington, DC: U.S. Patent and Trademark Office.
35. Reinhart, K. F., Eckardt, M., Herrmann, I., Feyh, A., & Freund, F. (2009). Low-cost Approach for Far-Infrared Sensor Arrays for Hot-Spot Detection in Automotive Night Vision Systems. In *Advanced Microsystems for Automotive Applications 2009* (pp. 397-408). Springer Berlin Heidelberg.
36. http://www.adose-eu.org/
37. http://www.autolivnightvision.com/technology/
38. http://www.bmw.com/com/en/insights/technology/technology_guide/articles/mm_bmw_night_vision.html
39. Pinchon, N., Ibn-Khedher, M., Cassignol, O., Nicolas, A., Bernardin, F., Leduc, P., ... & Julien, G. (2016, October). All-weather vision for automotive safety: which spectral band?. In *SIA Vision 2016-International Conference Night Drive Tests and Exhibition* (p. 7p). Société des Ingénieurs de l'Automobile-SIA.
40. http://image-sensors-world.blogspot.fr/2015/09/2016-bmw-7-series-gesture-control.html
41. Larson, E., Cohn, G., Gupta, S., Ren, X., Harrison, B., Fox, D., & Patel, S. (2011, May). HeatWave: thermal imaging for surface user interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2565-2574). ACM.
42. Abdelrahman, Y., Sahami Shirazi, A., Henze, N., & Schmidt, A. (2015, April). Investigation of material properties for thermal imaging-based interaction. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 15-18). ACM.
43. Yon, J. J., Dumont, G., Goudon, V., Becker, S., Arnaud, A., Cortial, S., & Tisse, C. L. (2014, June). Latest improvements in microbolometer thin film packaging: paving the way for low-cost consumer applications. In *SPIE Defense+ Security* (pp. 90701N-90701N). International Society for Optics and Photonics.
44. Guettari, T. (2014). *Détection de la présence humaine et évaluation de la qualité du sommeil en établissement d'hébergement pour personnes âgées dépendantes (EHPAD)* (Doctoral dissertation, Evry, Institut national des télécommunications).
45. VanSomeren, E. J. (2000). More than a marker: interaction between the circadian regulation of temperature and sleep, age-related changes, and treatment possibilities. *Chronobiology international*, *17*(3), 313-354.
46. Lack, L., & Gradisar, M. (2002). Acute finger temperature changes preceding sleep onsets over a 45-h period. *Journal of sleep research*, *11*(4), 275-282.
47. Nozawa, A., & Tacano, M. (2009). Correlation analysis on alpha attenuation and nasal skin temperature. *Journal of Statistical Mechanics: Theory and Experiment*, *2009*(01), P01007.
48. Kräuchi, K., Cajochen, C., Werth, E., & Wirz-Justice, A. (2000). Functional link between distal vasodilation and sleep-onset latency?. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *278*(3), R741-R748.
49. Sun, N., Garbey, M., Merla, A., & Pavlidis, I. (2005, June). Imaging the cardiovascular pulse. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 2, pp. 416-421). IEEE.

50. Liang, W. C., Yuan, J., Sun, D. C., & Lin, M. H. (2009). Changes in physiological parameters induced by indoor simulated driving: Effect of lower body exercise at mid-term break. *Sensors*, *9*(9), 6913-6933.
51. Murthy, R., & Pavlidis, I. (2006). Noncontact measurement of breathing function. *IEEE Engineering in Medicine and Biology Magazine*, *25*(3), 57-67.
52. AL-Khalidi, F. Q., Saatchi, R., Burke, D., Elphick, H., & Tan, S. (2011). Respiration rate monitoring methods: A review. *Pediatric pulmonology*, *46*(6), 523-529.
53. Stemberger, J., Allison, R. S., & Schnell, T. (2010, May). Thermal imaging as a way to classify cognitive workload. In *Computer and Robot Vision (CRV), 2010 Canadian Conference on* (pp. 231-238). IEEE.
54. Puri, C., Olson, L., Pavlidis, I., Levine, J., & Starren, J. (2005, April). StressCam: non-contact measurement of users' emotional states through thermal imaging. In *CHI'05 extended abstracts on Human factors in computing systems* (pp. 1725-1728). ACM.
55. Wesley, A., Shastri, D., & Pavlidis, I. (2010, April). A novel method to monitor driver's distractions. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 4273-4278). ACM.
56. Zhou, Y., Tsiamyrtzis, P., Lindner, P., Timofeyev, I., & Pavlidis, I. (2013). Spatiotemporal smoothing as a basis for facial tissue tracking in thermal imaging. *IEEE Transactions on Biomedical Engineering*, *60*(5), 1280-1289.
57. Pavlidis, I., Dcosta, M., Taamneh, S., Manser, M., Ferris, T., Wunderlich, R., ... & Tsiamyrtzis, P. (2016). Dissecting Driver Behaviors Under Cognitive, Emotional, Sensorimotor, and Mixed Stressors. *Scientific reports*, *6*.
58. Murphy-Chutorian, E., & Trivedi, M. M. (2009). Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *31*(4), 607-626.
59. Beymer, D. J., "Face recognition under varying pose", IEEE Computer Vision and Pattern Recognition, 756-761 (1994).
60. Murphy-Chutorian, E., & Trivedi, M. M. (2010). Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness. *IEEE Transactions on intelligent transportation systems*, *11*(2), 300-311.
61. Morency, L. P., Whitehill, J., & Movellan, J. (2010). Monocular head pose estimation using generalized adaptive view-based appearance model. *Image and Vision Computing*, *28*(5), 754-761.
62. Horprasert, T., Yacoob, Y., & Davis, L. S. (1996, October). Computing 3-d head orientation from a monocular image sequence. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on* (pp. 242-247). IEEE.
63. Tawari, A., Martin, S., & Trivedi, M. M. (2014). Continuous head movement estimator for driver assistance: Issues, algorithms, and on-road evaluations. *IEEE Transactions on Intelligent Transportation Systems*, *15*(2), 818-830.
64. Lepetit, V., & Fua, P. (2005). Monocular model-based 3d tracking of rigid objects: A survey. *Foundations and Trends® in Computer Graphics and Vision*, *1*(1), 1-89.

65. Wang, H., Davoine, F., Lepetit, V., Chaillou, C., & Pan, C. (2012). 3-d head tracking via invariant keypoint learning. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(8), 1113-1126.
66. Ström, J. (2002). Model-based real-time head tracking. *EURASIP Journal on Advances in Signal Processing*, *2002*(10), 873945.
67. Vacchetti, L., Lepetit, V., & Fua, P. (2004). Stable real-time 3d tracking using online and offline information. *IEEE transactions on pattern analysis and machine intelligence*, *26*(10), 1385-1391.
68. Wu, S., Jiang, L., Xie, S., & Yeo, A. C. (2006). A robust method for detecting facial orientation in infrared images. *Pattern recognition*, *39*(2), 303-309.
69. Yu, X., Chua, W. K., Dong, L., Hoe, K. E., & Li, L. (2010, May). Head pose estimation in thermal images for human and robot interaction. In *Industrial Mechatronics and Automation (ICIMA), 2010 2nd International Conference on* (Vol. 2, pp. 698-701). IEEE.
70. Kato, T., Fujii, T., & Tanimoto, M. (2004, June). Detection of driver's posture in the car by using far infrared camera. In *Intelligent Vehicles Symposium, 2004 IEEE* (pp. 339-344). IEEE.
71. Morris, N. J., Avidan, S., Matusik, W., & Pfister, H. (2007, June). Statistics of infrared images. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (pp. 1-7). IEEE.
72. Minassian, C., Tissot, J. L., Vilain, M., Legras, O., Tinnes, S., Fieque, B., ... & Robert, P. (2008, April). Uncooled amorphous silicon TEC-less 1/4 VGA IRFPA with 25 µm pixel-pitch for high volume applications. In *SPIE Defense and Security Symposium* (pp. 69401Z-69401Z). International Society for Optics and Photonics.
73. Maroto, O., Diez-Merino, L., Carbonell, J., Tomàs, A., Reyes, M., Joven-Alvarez, E., ... & Rodríguez-Frías, M. D. (2014, July). Design of the Front End Electronics for the Infrared Camera of JEM-EUSO, and manufacturing and verification of the prototype model. In *SPIE Astronomical Telescopes+ Instrumentation* (pp. 915424-915424). International Society for Optics and Photonics.
74. Pittaluga, F., Zivkovic, A., & Koppal, S. J. (2016, May). Sensor-level privacy for thermal cameras. In *Computational Photography (ICCP), 2016 IEEE International Conference on* (pp. 1-12). IEEE.
75. Xiu-Bao, S., Qian, C., Guo-Hua, G., & Ning, L. (2010). Research on the response model of microbolometer. *Chinese Physics B*, *19*(10), 108702.
76. Lee, J., & Kyung, C. M. (2013, June). Characterization of non-uniformity and bias-heating for uncooled bolometer FPA detectors using simulator. In *Proc. SPIE* (Vol. 8706, pp. 87060X-1).
77. Kim, G., & Ko, H. (2015, November). Behavioral modeling and experimental validation of uncooled microbolometer. In *SENSORS, 2015 IEEE* (pp. 1-3). IEEE.
78. Budzier, H., & Gerlach, G. (2011). *Thermal infrared sensors: theory, optimisation and practice*. John Wiley & Sons.
79. http://flir.custhelp.com/app/answers/detail/a_id/128/~/how-is-nedt-measured%3F
80. Guérineau, N., Haidar, R., Bernhardt, S., Ribet-Mohamed, I., & Caes, M. (2007). Caractérisations électro-optiques des détecteurs plans focaux IR. *Techniques de l'ingénieur. Mesures et contrôle*, (R6460).

81. Cao, Y., & Tisse, C. L. (2013). Shutterless solution for simultaneous focal plane array temperature estimation and nonuniformity correction in uncooled long-wave infrared camera. *Applied optics*, *52*(25), 6266-6271.
82. Bieszczad, G., Orżanowski, T., Sosnowski, T., & Kastek, M. (2009, September). Method of detectors offset correction in thermovision camera with uncooled microbolometric focal plane array. In *SPIE Europe Security+ Defence* (pp. 74810O-74810O). International Society for Optics and Photonics.
83. Perry, D. L., & Dereniak, E. L. (1993). Linear theory of nonuniformity correction in infrared staring sensors. *Optical Engineering*, *32*(8), 1854-1859.
84. Ariyaratnam, S., & Rood, J. P. (1990). Measurement of facial skin temperature. *Journal of dentistry*, *18*(5), 250-253.
85. Liang, K., Yang, C., Peng, L., & Zhou, B. (2017). Nonuniformity correction based on focal plane array temperature in uncooled long-wave infrared cameras without a shutter. *Applied Optics*, *56*(4), 884-889.
86. Säuberlicha, T., Paprotha, C., & Helbertb, J. (2010, August). MERTIS–Shutterless Background Signal Removal. In *SPIE Optical Engineering+ Applications* (pp. 78080L-78080L). International Society for Optics and Photonics.
87. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), 91-110.
88. Olbrycht, R., Wiecek, B., & Swiatczak, T. (2009, June). Shutterless method for gain nonuniformity correction of microbolometer detectors. In *Mixed Design of Integrated Circuits & Systems, 2009. MIXDES'09. MIXDES-16th International Conference* (pp. 378-380). IEEE.
89. Olbrycht, R., Więcek, B., & De Mey, G. (2012). Thermal drift compensation method for microbolometer thermal cameras. *Applied optics*, *51*(11), 1788-1794.
90. Harris, J. G., & Chiang, Y. M. (1997, August). Nonuniformity correction using the constant-statistics constraint: analog and digital implementations. In *AeroSense'97* (pp. 895-905). International Society for Optics and Photonics.
91. Narayanan, B., Hardie, R. C., & Muse, R. A. (2005). Scene-based nonuniformity correction technique that exploits knowledge of the focal-plane array readout architecture. *Applied optics*, *44*(17), 3482-3491.
92. R. C. Hardie, M. M. Hayat, E. Armstrong, et B. Yasuda, « Scene-Based Nonuniformity Correction with Video Sequences and Registration », *Appl. Opt.*, vol. 39, n° 8, p. 1241‑1250, mars 2000.
93. Buades, A., Coll, B., & Morel, J. M. (2005, June). A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 2, pp. 60-65). IEEE.
94. Cao, Y., & Li, Y. (2015). Strip non-uniformity correction in uncooled long-wave infrared focal plane array based on noise source characterization. *Optics Communications*, *339*, 236-242.
95. Tendero, Y., Gilles, J., Landeau, S., & Morel, J. M. (2010, October). Efficient single image non-uniformity correction algorithm. In *Security+ Defence* (pp. 78340E-78340E). International Society for Optics and Photonics.

96. Tendero, Y., & Gilles, J. (2012, May). ADMIRE: a locally adaptive single-image, non-uniformity correction and denoising algorithm: application to uncooled IR camera. In *SPIE Defense, Security, and Sensing* (pp. 83531O-83531O). International Society for Optics and Photonics.
97. Wang, S. P. (2016). Stripe noise removal for infrared image by minimizing difference between columns. *Infrared Physics & Technology*, *77*, 58-64.
98. Zhao, J., Zhou, Q., Chen, Y., Liu, T., Feng, H., Xu, Z., & Li, Q. (2013). Single image stripe nonuniformity correction with gradient-constrained optimization model for infrared focal plane arrays. *Optics Communications*, *296*, 47-52.
99. Delon, J. (2004). Midway image equalization. *Journal of Mathematical Imaging and Vision*, *21*(2), 119-134.
100. Saragaglia, A., & Durand, A. (2015). *U.S. Patent Application No. 14/695,539*.
101. Bai, J., Ma, Y., Li, J., Fan, F., & Wang, H. (2011). Novel averaging window filter for SIFT in infrared face recognition. *Chinese Optics Letters*, *9*(8), 081002-081002.
102. Nugent, P. W., Shaw, J. A., & Pust, N. J. (2013). Correcting for focal-plane-array temperature dependence in microbolometer infrared cameras lacking thermal stabilization. *Optical Engineering*, *52*(6), 061304-061304.
103. Nugent, P. W., Shaw, J. A., & Pust, N. J. (2014). Radiometric calibration of infrared imagers using an internal shutter as an equivalent external blackbody. *Optical Engineering*, *53*(12), 123106-123106.
104. Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(11), 1330-1334.
105. Heikkila, J., & Silvén, O. (1997, June). A four-step camera calibration procedure with implicit image correction. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on* (pp. 1106-1112). IEEE.
106. Yu, S., Kim, J., & Lee, S. (2012). Thermal 3D modeling system based on 3-view geometry. *Optics Communications*, *285*(24), 5019-5028.
107. Hartley, R. I. (1997). In defense of the eight-point algorithm. *IEEE Transactions on pattern analysis and machine intelligence*, *19*(6), 580-593.
108. Nistér, D. (2004). An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, *26*(6), 756-770.
109. http://www.3dsystems.com/shop/scanners
110. http://www.artcreaxtion.com/3d-depth-sensor/apple-primesense
111. Zollhöfer, M., Martinek, M., Greiner, G., Stamminger, M., & Süßmuth, J. (2011). Automatic reconstruction of personalized avatars from 3D face scans. *Computer Animation and Virtual Worlds*, *22*(2-3), 195-202.
112. Tang, L. A., & Huang, T. S. (1996, September). Automatic construction of 3D human face models based on 2D images. In *Image Processing, 1996. Proceedings., International Conference on* (Vol. 3, pp. 467-470). IEEE.
113. http://www.icg.isy.liu.se/candide/

114. Strom, J., Jebara, T., Basu, S., & Pentland, A. (1999). Real time tracking and modeling of faces: An ekf-based analysis by synthesis approach. In *Modelling People, 1999. Proceedings. IEEE International Workshop on* (pp. 55-61). IEEE.
115. http://visagetechnologies.com/
116. Zhu, Z., & Ji, Q. (2004, June). Real time 3d face pose tracking from an uncalibrated camera. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on* (pp. 73-73). IEEE.
117. Jang, J. S., & Kanade, T. (2008, September). Robust 3D head tracking by online feature registration. In *8th IEEE Int. Conf. on Automatic Face and Gesture Recognition*.
118. Basu, S., Essa, I., & Pentland, A. (1996, August). Motion regularization for model-based head tracking. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on* (Vol. 3, pp. 611-616). IEEE.
119. Ceylan, M., Henriques, D. Y. P., Tweed, D. B., & Crawford, J. D. (2000). Task-dependent constraints in motor control: pinhole goggles make the head move like an eye. *The Journal of Neuroscience*, *20*(7), 2719-2730.
120. Kunin, M., Osaki, Y., Cohen, B., & Raphan, T. (2007). Rotation axes of the head during positioning, head shaking, and locomotion. *Journal of neurophysiology*, *98*(5), 3095-3108.
121. Moore, S. T., Hirasaki, E., Raphan, T., & Cohen, B. (2005). Instantaneous rotation axes during active head movements. *Journal of Vestibular Research*, *15*(2), 73-80.
122. Medendorp, W. P., Melis, B. J. M., Gielen, C. C. A. M., & Van Gisbergen, J. A. M. (1998). Off-centric rotation axes in natural head movements: implications for vestibular reafference and kinematic redundancy. *Journal of neurophysiology*, *79*(4), 2025-2039.
123. Murphy-Chutorian, E., & Trivedi, M. M. (2008, June). Hyhope: Hybrid head orientation and position estimation for vision-based driver head tracking. In *Intelligent Vehicles Symposium, 2008 IEEE* (pp. 512-517). IEEE.
124. Morency, L. P., Whitehill, J., & Movellan, J. (2010). Monocular head pose estimation using generalized adaptive view-based appearance model. *Image and Vision Computing*, *28*(5), 754-761.
125. Wang, J. G., & Sung, E. (2007). EM enhancement of 3D head pose estimated by point at infinity. *Image and Vision Computing*, *25*(12), 1864-1874.
126. InterSense Inc. *Inertia Cube3 Manual.* http://www.intersense.com
127. Mekyska, J., Espinosa-Duró, V., & Faundez-Zanuy, M. (2010, October). Face segmentation: A comparison between visible and thermal images. In *Security Technology (ICCST), 2010 IEEE International Carnahan Conference on* (pp. 185-189). IEEE.
128. Kittler, J., & Illingworth, J. (1985). On threshold selection using clustering criteria. *IEEE transactions on systems, man, and cybernetics*, (5), 652-655.
129. Socolinsky, D. A., Wolff, L. B., Neuheisel, J. D., & Eveland, C. K. (2001). Illumination invariant face recognition using thermal infrared imagery. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (Vol. 1, pp. I-I). IEEE.
130. Krotosky, S. J., Cheng, S. Y., & Trivedi, M. M. (2004, October). Face detection and head tracking using stereo and thermal infrared cameras for" smart" airbags: a comparative analysis. In *Intelligent*

Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on (pp. 17-22). IEEE.

131. Tu, J., Huang, T., & Tao, H. (2006, April). Accurate head pose tracking in low resolution video. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on* (pp. 573-578). IEEE.
132. Soulez, F., Denis, L., Fournier, C., Thiébaut, É., & Goepfert, C. (2007). Inverse-problem approach for particle digital holography: accurate location based on local optimization. *JOSA A*, *24*(4), 1164-1171.
133. Fisher, R. B., & Naidu, D. K. (1996). A comparison of algorithms for subpixel peak detection. In *Image technology* (pp. 385-404). Springer Berlin Heidelberg.
134. Lewis, J. P. (1995, May). Fast normalized cross-correlation. In *Vision interface* (Vol. 10, No. 1, pp. 120-123).
135. Gourier, N., Maisonnasse, J., Hall, D., & Crowley, J. L. (2006, April). Head pose estimation on low resolution images. In *International Evaluation Workshop on Classification of Events, Activities and Relationships* (pp. 270-280). Springer Berlin Heidelberg.
136. McKenna, S. J., & Gong, S. (1998). Real-time face pose estimation. *Real-Time Imaging*, *4*(5), 333-347.
137. Srinivasan, S., & Boyer, K. L. (2002). Head pose estimation using view based eigenspaces. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on* (Vol. 4, pp. 302-305). IEEE.
138. Wu, J., & Trivedi, M. M. (2008). A two-stage head pose estimation framework and evaluation. *Pattern Recognition*, *41*(3), 1138-1158.
139. Johansson, J., Solli, M., & Maki, A. (2016, October). An Evaluation of Local Feature Detectors and Descriptors for Infrared Images. In *Computer Vision–ECCV 2016 Workshops* (pp. 711-723). Springer International Publishing.
140. Ricaurte, P., Chilán, C., Aguilera-Carrasco, C. A., Vintimilla, B. X., & Sappa, A. D. (2014). Feature point descriptors: Infrared and visible spectra. *Sensors*, *14*(2), 3690-3701.
141. Tuytelaars, T., & Mikolajczyk, K. (2008). Local invariant feature detectors: a survey. *Foundations and trends® in computer graphics and vision*, *3*(3), 177-280.
142. Harris, C., & Stephens, M. (1988, August). A combined corner and edge detector. In *Alvey vision conference* (Vol. 15, p. 50).
143. Moravec, H. P. (1980). *Obstacle avoidance and navigation in the real world by a seeing robot rover* (No. STAN-CS-80-813). Standorf Univ. CA Dept. of computer science.
144. Koenderink, J. J., & van Doorn, A. J. (1987). Representation of local geometry in the visual system. *Biological cybernetics*, *55*(6), 367-375.
145. Kitchen, L., & Rosenfeld, A. (1980). Gray-level corner detection. *Pattern recognition letters*, 1:95-102, 1982.
146. Witkin, A. (1984, March). Scale-space filtering: A new approach to multi-scale description. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.* (Vol. 9, pp. 150-153). IEEE.
147. Koenderink, J. J. (1984). The structure of images. *Biological cybernetics*, *50*(5), 363-370.

148. Lindeberg, T. (1994). Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, *21*(1-2), 225-270.
149. Florack, L. M. J. (1993). *The syntactical structure of scalar images*. Universiteit Utrecht, Faculteit Geneeskunde.
150. Lindeberg, T. (1998). Feature detection with automatic scale selection. *International journal of computer vision*, *30*(2), 79-116.
151. Mikolajczyk, K., Detection of local features invariant to affine transformations, *Doctoral dissertation, Institut National Polytechnique de Grenoble* (2002).
152. Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on* (Vol. 2, pp. 1150-1157). Ieee.
153. https://fr.mathworks.com/matlabcentral/fileexchange/17894-keypoint-extraction
154. http://www.vlfeat.org/overview/sift.html
155. Schmid, C., Mohr, R., & Bauckhage, C. (1998, January). Comparing and evaluating interest points. In *Computer Vision, 1998. Sixth International Conference on* (pp. 230-235). IEEE.
156. Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886-893). IEEE.
157. Grunert, J. A. (1841). Das pothenotische problem in erweiterter gestalt nebst über seine anwendungen in der geodäsie. *Grunerts archiv für mathematik und physik*, *1*, 238-248.
158. Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381-395.
159. Horn, B. K., Hilden, H. M., & Negahdaripour, S. (1988). Closed-form solution of absolute orientation using orthonormal matrices. *JOSA A*, *5*(7), 1127-1135.
160. Quan, L., & Lan, Z. (1999). Linear n-point camera pose determination. *IEEE Transactions on pattern analysis and machine intelligence*, *21*(8), 774-780.
161. Umeyama, S. (1991). Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, *13*(4), 376-380.
162. Horn, B. K. (1987). Closed-form solution of absolute orientation using unit quaternions. *JOSA A*, *4*(4), 629-642.
163. Haralick, B. M., Lee, C. N., Ottenberg, K., & Nölle, M. (1994). Review and analysis of solutions of the three point perspective pose estimation problem. *International journal of computer vision*, *13*(3), 331-356.
164. Li, S., & Xu, C. (2011). A stable direct solution of perspective-three-point problem. *International journal of pattern recognition and artificial intelligence*, *25*(05), 627-642.
165. Li, S., Xu, C., & Xie, M. (2012). A robust O (n) solution to the perspective-n-point problem. *IEEE transactions on pattern analysis and machine intelligence*, *34*(7), 1444-1450.
166. Lepetit, V., Moreno-Noguer, F., & Fua, P. (2009). Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, *81*(2), 155-166.

167. Ferraz, L., Binefa, X., & Moreno-Noguer, F. (2014). Very fast solution to the PnP problem with algebraic outlier rejection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 501-508).
168. Bai, J., Ma, Y., Li, J., Li, H., Fang, Y., Wang, R., & Wang, H. (2014). Good match exploration for thermal infrared face recognition based on YWF-SIFT with multi-scale fusion. *Infrared Physics & Technology*, *67*, 91-97.
169. Hu, S., Choi, J., Chan, A. L., & Schwartz, W. R. (2015). Thermal-to-visible face recognition using partial least squares. *JOSA A*, *32*(3), 431-442.
170. Berry, S., Bozler, C. O., Reich, R. K., Clark, H. R., Bos, P., Finnemeyer, V., & McGinty, C. (2016). A Scalable Fabrication Process for Liquid Crystal-Based Uncooled Thermal Imagers. Journal of Microelectromechanical Systems, 25(3), 479-488.
171. Laurent, L., Yon, J. J., Moulet, J. S., Imperinetti, P., & Duraffourg, L. (2016, September). Low noise torsional mechanical resonator for 12μm microbolometers. In SPIE Optical Engineering+ Applications (pp. 997407-997407). International Society for Optics and Photonics.