Champ visuel augmenté pour l'exploration vidéo de la rétine

Alexandre Guerre

UNIVERSITE BIOLOGIE BRETAGNE SANTE LOIRE

UBO
Université de Bretagne Occidentale

THESE DE DOCTORAT DE

L'UNIVERSITE
DE BRETAGNE OCCIDENTALE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : *Analyse et Traitement de l'Information et des Images Médicales*

Par
Alexandre GUERRE

Champ visuel augmenté pour exploration vidéo de la rétine.

Thèse présentée et soutenue à Brest, le 20/12/2019
Unité de recherche : UMR1101 Inserm, LaTIM

Rapporteurs avant soutenance :

Sandrine VOROS PhD – HDR – CR Inserm, Pavillon Taillefer, Laboratoire TIMC-IMAG
Marc MURAINE PUPH d'ophtalmologie, CHU de Rouen, Service d'ophtalmologie

Composition du Jury :

Président : Marc MURAINE PUPH d'ophtalmologie, CHU de Rouen, Service d'ophtalmologie

Examinateur : Sandrine VOROS PhD – HDR – CR Inserm, Pavillon Taillefer, Laboratoire TIMC-IMAG

Dir. de thèse :
Gwenolé QUELLEC PhD – HDR – CR Inserm, UMR1101 Inserm, LaTIM

Co-dir. de thèse :
Béatrice COCHENER PUPH d'ophtalmologie, CHU de Brest, Service d'ophtalmologie

Invité(s)
Mathieu LAMARD PhD – IR UBO, UMR1101 Inserm, LaTIM
Pierre-Henri CONZE PhD – MC IMTA, UMR1101 Inserm, LaTIM

Table des matières

Table des figures

Liste des tableaux

1 Introduction

Depuis Hippocrate la médecine n'a cessé d'évoluer. L'alliance de ce domaine avec la physique a permis la création de l'imagerie médicale à la fin du XIXème siècle. Dès lors, de nombreux systèmes ont été développés et perfectionnés pour proposer des images toujours plus précises de l'intérieur du corps humain.

Quelques décennies plus tard, dans la première moitié du XXème siècle, la lampe à fente et l'endoscope oculaire voient le jour. Ces deux outils permettent de visualiser l'œil et plus spécifiquement la rétine. C'est justement sur cette partie que nous allons nous focaliser. En effet, le but de cette thèse est d'apporter un meilleur confort aux ophtalmologues lors d'examens ou de chirurgies rétiniennes.

Sur les vidéos acquises à l'endoscope oculaire et, dans une moindre mesure, sur les vidéos acquises à la lampe à fente, nous constatons que le champ visuel proposé est assez faible. Il arrive même, lors de certaines interventions (pour des décollements de rétines par exemple), que les chirurgiens aient des difficultés à estimer quelles parties de la rétine ont déjà été traitées et lesquelles ne l'ont pas encore été. De plus, il n'est pas aisé d'estimer précisément les distances, et plus particulièrement la profondeur, avec de tels outils. Voilà pourquoi, en endoscopie, il arrive que le chirurgien sorte l'outil de l'œil sans pour autant le vouloir.

Pour l'ensemble de ces raisons nous proposons d'augmenter, en temps réel, le champ visuel des images acquises à la lampe à fente et à l'endoscope oculaire, en réalisant des cartes dynamiques de la rétine. Il n'existe pas, à notre connaissance, de travaux se penchant sur cette problématique dans l'état de l'art pour des acquisitions à l'endoscope oculaire.

Dans un premier temps nous orientons nos recherches vers la réalisation de mosaïques en deux dimensions de la rétine avant de rajouter l'information sur la profondeur. Pour ce faire nous cherchons à estimer précisément les déplacements entre deux images à travers plusieurs méthodes.

Nous nous penchons notamment sur une méthode d'estimation dense de déplacements

en calculant le flux optique entre des paires d'images. Pour résumer, cette méthode attribue à chaque pixel un polynôme quadratique correspondant au voisinage proche de celui-ci puis cherche à mettre en correspondance les polynômes entre les images.

Nous étudions également des méthodes parcimonieuses détectant automatiquement un certain nombre de points d'intérêt dans les images. Une fois les points détectés un algorithme se charge de les faire correspondre. A partir des points appariés, un autre algorithme estime un modèle mathématique du déplacement entre les deux images.

Une troisième approche utilise une méthode initialement utilisée pour l'encodage de vidéos. Elle consiste à diviser les images en blocs et à chercher les blocs les plus proches dans les autres images.

Pour finir, des méthodes utilisant l'apprentissage profond sont étudiées. Il s'agit de réseaux de neurones convolutifs spécialisés dans l'estimation de déplacement. L'estimation de la profondeur se fait également par l'intermédiaire d'un réseau de neurone convolutif.

Pour réaliser nos tests et entraîner les réseaux de neurones, plusieurs bases de données ayant pour fond la rétine sont construites. Une première est constituée à partir des vidéos acquises à l'endoscope oculaire, une seconde à partir des vidéos acquises à la lampe à fente.

Enfin, deux autres bases illustrent des déplacements artificiels constitués à partir de paires d'images extraites de fonds d'œil. Ces déplacements artificiels étant générés par nos soins, nous disposons donc de la vérité terrain du déplacement d'une image à l'autre. A la différence des deux premières bases de données, celles-ci peuvent donc être considérées comme annotées.

1.1 Plan

Dans ce manuscrit, le plan se déroule comme suit :

Le chapitre 2 introduit le contexte de cette thèse et propose un état de l'art des différentes méthodes développées pour l'estimation de déplacements entre images.

Le chapitre 3 présente brièvement les réseaux de neurones avant de se concentrer sur les réseaux de neurones convolutifs et leurs structures.

Le chapitre 4 retrace l'historique et les évolutions des deux dispositifs d'acquisitions d'images étudiés dans cette thèse, à savoir : la lampe à fente et l'endoscope oculaire. Il détaille également les différentes bases de données utilisées et développées dans le cadre de cette étude.

Le chapitre 5 traite des différentes méthodes utilisées traditionnellement dans la littérature pour l'estimation de déplacements entre deux images. De plus, il présente les résultats de ces méthodes sur nos bases de données.

Le chapitre 6 aborde les méthodes utilisant des réseaux de neurones convolutifs pour l'estimation de déplacements et présente leurs résultats sur nos différentes bases de données.

Le chapitre 7 conclut le manuscrit en apportant un bilan du travail réalisé et en proposant des pistes pour de possibles améliorations des résultats.

2

Etat de l'art pour l'augmentation du champ visuel

Le domaine de l'augmentation du champ visuel est très vaste. Son application la plus connue du grand public est sans doute celle du panorama en photographie. En effet, pour réaliser des panoramas [3], une série d'images différentes sont acquises avec un certain taux de recouvrement entre elles. C'est grâce à ce recouvrement que des zones communes vont pouvoir être déterminées d'une image à l'autre par un algorithme. Une fois les zones communes identifiées, il est possible d'estimer le déplacement entre les images et de les assembler pour former des panoramas.

On retrouve également des applications dans le domaine de l'astronomie [4] où l'objectif est d'arriver à détecter automatiquement certains groupes d'étoiles pour constituer des cartographies et des atlas à partir d'images de télescope. Dans la direction inverse, des images acquises par satellites ou drones [5], [6], [7], [8], servent à cartographier des territoires terrestres pour déterminer la surface de certaines régions ou encore dissocier automatiquement plusieurs zones.

Des applications dans le domaine médical existent aussi. On retrouve par exemple des mosaïques de corps entiers réalisées à base de volumes acquis par IRM (Imagerie par Résonance Magnétique) [9]. Une application plus commune est la cartographie du système digestif grâce à des capsules endoscopiques à avaler par le patient [10]. Ces dispositifs permettent de déterminer l'ensemble des zones malignes et de les localiser précisément avant l'opération. Il peut arriver que les capsules soient équipées d'un système de localisation pour savoir précisément où se trouvent ces zones sur le corps du patient.

Pendant la chirurgie une augmentation du champ visuel peut aussi être utile. Ainsi, plusieurs études ont eu pour but de cartographier à la volée le système digestif pendant des laparoscopies [11] ou, plus précisément, de cartographier et délimiter des

zones malignes [12]. Cette mise en place a également permis une aide au diagnostic pour les chirurgiens ainsi que des interventions chirurgicales plus précises. Le but est le même dans les cas de chirurgies du placenta sous endoscope lors de grossesses compliquées [13],[14].

Enfin, dans le domaine oculaire, ces pratiques ont déjà été mises en places pour des chirurgies de la rétine sous microscope [15]. On peut aussi retrouver des constructions de mosaïques d'images dans le cas d'examens réalisés à la lampe à fente [16], [17], [18]. En revanche, il ne semble pas exister, dans la littérature actuelle, d'application dédiée à la réalisation de mosaïques à partir d'images issues d'endoscope oculaire.

La suite de ce chapitre s'articule en deux parties principales. La première détaille les méthodes qualifiées de "classiques" pour la réalisation de mosaïque d'images et la seconde se concentre sur des méthodes utilisant des concepts plus récents basés sur l'apprentissage profond. Dans cette sous-partie, deux méthodes sont mises en avant : la méthode FlowNetS et la méthode SFM Learner). Il s'agit des méthodes principalement utilisées dans le cadre de cette thèse.

2.1 Méthodes classiques

On peut regrouper les méthodes classiques en trois groupes principaux. Le premier porte sur l'étude du flux optique. Cette notion est développée dans la partie suivante. On la retrouve peu dans la partie médicale de l'estimation des déplacements pour la construction de mosaïques, mais elle constitue tout de même une part importante de la littérature de l'estimation des déplacements de manière générale. Les méthodes du second groupe sont elles aussi assez peu utilisées à des fins médicales, mais sont une partie non négligeable dans l'état de l'art de l'estimation de déplacements. Enfin, les méthodes du troisième groupe sont globalement plus récentes et leurs usages dans le domaine médical est plus commun.

2.1.1 Les méthodes utilisant le flux optique

Le flux optique est une représentation du mouvement des objets dans une vidéo. Par extension, il peut désigner l'ensemble des déplacements entre deux images. Ce concept date du début des années 1950 [19]. Deux méthodes datant des années 1980 servent encore aujourd'hui de référence dans la littérature. Il s'agit des méthodes de Lucas-Kanade [20] et de Horn-Schunck [21]. La première est une méthode dite locale puisque pour calculer le flux optique, elle se base sur un voisinage proche du pixel. A l'inverse, la méthode de Horn Schunck est une méthode globale qui va rechercher une certaine homogénéité dans le flux des images. Depuis, plusieurs autres méthodes ont été développées. Celle qui semble être la plus performante dans la littérature est la méthode de Farnebäck [22].

Cette dernière a d'ailleurs été adaptée à de l'estimation de mouvement sur des séquences 3D [23]. Le concept de la méthode de Farnebäck est basée sur la décomposition polynomiale des signaux, dans son cas des images 2D. Dans [23], Danudibroto et al. s'en inspirent et réalisent des cartes 3D de volumes d'échocardiographie.

La structure globale de leur méthode est assez classique. On trouve dans un premier temps une étape de pré-traitement des données, composée de la mise en place d'un masque et de filtrages pour mettre en évidence les structures importantes. Ils appliquent ensuite leur adaptation du calcul du flux optique de Farnebäck sur leurs données. Enfin, ils calculent un déplacement global entre les paires d'images volumiques et mettent en place la mosaïque 3D.

Chiba et al. ont développé dans [24] une méthode généraliste de création de mosaïque d'images (2D) basée sur la méthode de Lucas-Kanade. Ils pré-traitent les images en les sous échantillonnant pour rendre les calculs plus rapides. Ils estiment grossièrement les zones de recouvrement entre les paires d'images en cherchant à maximiser un score de corrélation croisée. Par la suite, ils n'estiment les flux optique, que sur des patchs provenant des zones présentant les scores les plus élevées à l'étape de corrélation croisée. En supposant que les mouvements à trouver sont de forme homographique, ils résolvent des systèmes d'équations avec pour inconnues les paramètres de l'homographie et pour variables les flux optique estimés pour des patchs.

Birchfield et al. choisissent de combiner les méthodes Lucas-Kanade et Horn-Schunck [25]. Ils aboutissent à une méthode hybride qui va apporter l'information des flux globaux (fournie par Horn-Schunck) des paires d'images aux flux locaux (fournis pas Lucas-Kanade) préalablement calculés. Quelques années auparavant, Bruhn et al. [26] ont eux aussi tenté de combiner les informations de ces deux méthodes. Ces méthodes se placent dans des cas où l'éclairage entre les prises de vue reste constant de manière à conserver au maximum les intensités des pixels communs d'une scène à l'autre. C'est d'ailleurs la principale limite de ce type de méthode.

Trinh et al. [27] tentent de solutionner le problème de variation d'éclairage en introduisant un indice de confiance dans le calcul du flux optique. Cet indice est calculé dans l'étape de pré-traitement en changeant d'espace colorimétrique. L'espace couramment utilisé en traitement d'image est le RVB ou RGB pour Rouge, Vert, Bleu (Red Green Blue en anglais). L'espace utilisé dans le travail de pré-traitement de l'article est XYZ où X représente les composantes vertes et rouges de l'image, Y la luminance de l'image et Z la composante bleue. Par la suite, la méthode qu'ils utilisent pour estimer le déplacement n'est pas une méthode d'estimation de flux optique classique. Elle est inspirée de la méthode de Hu et al. [28], elle-même inspirée de la méthode de Barnes et al. [29] qui consiste à découper les images en patchs et à chercher, pour chaque patch d'une image, quels sont les patchs les plus proches dans les images voisines. Ce principe ressemble d'ailleurs fortement à celui détaillé dans la partie suivante.

2.1.2 Les méthodes utilisant le block matching

Le block matching, que l'on pourrait traduire par appariement de blocs, est un concept qui date des années 80 [30] et qui a été mis en place pour la compression de vidéos. Il va puiser ses principes dans les travaux des années 70 de Limb et al. et Rocca et al., principalement dans les articles [31] et [32] sur l'estimation de déplacements d'images télévisuelles.

Comme pour les méthodes précédentes, l'estimation se fait entre deux images d'une paire. Une image est définie comme image de référence et l'autre comme image cible. L'image de référence est divisée en blocs de référence. L'algorithme consiste à chercher, pour chaque bloc de référence, le bloc candidat le plus proche en matière d'intensité, dans l'image cible [33]. La distance séparant les deux blocs correspond au déplacement estimé pour les éléments qui constituent ces blocs. En fonction des méthodes, la taille des blocs peut varier, tout comme la taille du voisinage dans lequel se fait la recherche du bloc correspondant dans l'image cible.

L'article [34] répertorie d'ailleurs les principales méthodes de block matching existantes et utilisées pour les encodages des normes MPEG1 à MPEG4. On y découvre que la taille usuelle des blocs est de 16×16 pixels et que la zone de recherche explorée dans l'image cible est 23×23 pixels. Sachant qu'initialement on centre le bloc candidat dans la zone de recherche, les déplacements estimés peuvent donc avoir une amplitude maximale de 7 pixels dans chaque direction. Le déplacement retenu sera celui qui minimise une fonction de coût. Celle-ci peut être calculée de plusieurs manières, mais elles sont globalement proches. Parmi les plus classiques on trouve la somme des différences absolues, qui est la moins coûteuse en temps de calcul, la moyenne des différences absolues et la moyenne des erreurs quadratiques. Il existe également des fonctions de coût basées sur l'étude rapport signal sur bruit entre l'image de référence et l'image créée par le mouvement hypothétique.

Afin de minimiser le temps de calcul, plusieurs études ont aussi été faites en faisant varier la manière dont l'espace de recherche est parcouru. La technique la plus coûteuse est la recherche exhaustive et comme son nom l'indique, elle va parcourir tout l'espace de recherche avant de proposer la solution qui minimise la fonction de coût. Plus tard sont développées des méthodes itératives basées sur plusieurs itérations au cours desquelles des minimas locaux vont être calculés pour guider les itérations suivantes [35], [36]. Une amélioration consistant à tenir compte des déplacements des blocs précédents dans l'étude du déplacement du bloc actuel a aussi été mise en place [37]. Enfin, des questions se sont posées sur l'optimisation du voisinage à considérer à chaque itération pour chaque pixel d'un bloc. Ainsi, il est possible de trouver des voisinages carrés de taille 5×5 pixels pour certaines itérations et 3×3 pour d'autres. On trouve aussi des voisinages de tailles variables, mais avec des formes de losanges et non plus de carrés, ce sont les méthodes Diamond Search.

Les méthodes Diamond Search sont encore actuellement utilisées et améliorées. On la retrouve par exemple dans [38] qui est inspirée de [39] elle-même inspirée des algorithmes classiques de block matching. Il s'agit en fait d'une méthode itérative qui, à chaque itération peut faire varier la taille de la forme élémentaire utilisée (ici un losange) ainsi que l'amplitude du voisinage exploré. Ce procédé permet une exploration plus précise et efficace du voisinage, faisant ainsi converger la fonction de coût plus rapidement que les autres méthodes. La fonction de coût quant à elle, étudie la somme des différences absolues entre les intensités des pixels du bloc de référence et celles des pixels des blocs candidats considérés.

On retrouve assez peu le block matching dans des applications médicales. Il est toutefois évoqué dans [40] qui vise à faire des mosaïques avec des images microscopiques de colon de souris. Il intervient aussi dans des méthodes de créations de

mosaïques plus généralistes [41]. Enfin, on peut trouver des applications dans le domaine océanographique avec la création de mosaïques de fonds marins à partir d'images acquises au sonar [42], [43]. Dans ces articles, le block matching est respectivement combiné à des réseaux de neurones ou à de la détection de points d'intérêt. Les méthodes basées sur ces derniers font d'ailleurs l'objet de la partie suivante.

2.1.3 Les méthodes utilisant des points d'intérêt

Plusieurs méthodes de l'état de l'art ont pour objectif de détecter automatiquement des points d'intérêt sur deux images (ou plus). Ce concept a été introduit au début des années 1980 par Moravec [44]. Une fois les points détectés elles essayent de retrouver des points d'intérêt communs entre deux images et de les apparier. Enfin, avec les coordonnées de suffisamment de paires de points elles vont retrouver le mouvement global qui lie les deux images.

Les méthodes principalement utilisées pour la détection automatique de points d'intérêt font appel à des algorithmes basés sur la détection d'angle comme SIFT (Scale Invariant Feature Transform) [45], SURF (Speed Up Robust Feature) [46] ou encore FAST (Features from Accelerated Segment Test) [47]. L'algorithme SIFT, qui est le plus ancien des trois, convolue l'image à étudier avec une série de filtres gaussiens. Leurs noyaux étant de plus en plus gros, il vont permettre une mise en évidence des contours principaux de la scène. Cette méthode est efficace, mais plutôt lente voilà pourquoi des travaux ont été réalisés pour essayer d'être plus rapide. La seconde méthode est aujourd'hui la référence dans la littérature. Pour trouver les points d'intérêts, elle utilise les réponses des ondelettes de Haar sur les images. Comme son nom l'indique, il s'agit d'une méthode plus rapide que SIFT. La troisième méthode est caractérisée par sa grande rapidité et est encore plus rapide que SURF. En revanche, elle est très sensible au bruit, là où SURF est assez robuste. De plus, il faut que les images présentent un contraste élevé pour un fonctionnement optimal. Elle est donc rarement utilisée en imagerie médicale.

En ce qui concerne la recherche de correspondances entre les listes de points d'intérêt de deux images, la méthode étalon est la méthode RanSaC (Random Sample Consensus) [48]. Elle fournit des listes de probabilités d'appariements des points d'intérêts.

On retrouve par exemple l'association des algorithmes SURF et RanSaC dans la méthode développée dans [49], ou encore [13] qui réalise des mosaïques de placenta sous endoscopie. Cette méthode se divise en 5 étapes principales. La première étape est dédiée au calibrage de la caméra. Le calibrage sert à corriger les déformations de l'image. Une partie expliquant plus en détail le calibrage (et surtout l'auto-calibrage) est développée au chapitre 6. Pour réaliser cette opération, un dispositif accepté au bloc opératoire a été mis en place. Dans celui-ci le chirurgien doit placer l'endoscope réglé avec la mise au point qu'il utilisera pendant l'intervention. Une fois l'outil calibré, le chirurgien peut l'introduire dans le placenta et commencer à le scanner. C'est là que débute la seconde étape et que SURF et RanSaC sont utilisés. Les étapes 3 et 4 sont dédiées au calcul du déplacement entre les images. Enfin, l'étape 5 a pour but de construire la mosaïque et de l'homogénéiser pour qu'elle ne présente pas d'aberrations visuellement. On retrouve également SURF dans [50] et [51] pour

la réalisation de mosaïques d'images acquises au microscope.

Avec une problématique similaire à [13], Daga et al. ont choisi d'utiliser SIFT dans [14] pour cartographier le placenta. SIFT apparaît aussi pour la réalisation de mosaïques du tube digestif par capsule endoscopique par Maciura et al. [10]. Une fois les points d'intérêt détectés, ils utilisent un algorithme des K plus proches voisins, aussi appelé K-NN (K-Nearest Neighbors) pour les apparier. En utilisant la distance euclidienne et en prenant K=2, ils apparient 4 points pour chaque paire d'image. Comme ils recherchent des mouvements plans et sans changement de perspective, 4 points sont suffisants. Dans cette méthode il n'y a pas d'étape dédiée au calibrage, car la capsule est déjà pré-calibrée avant leur utilisation.

2.2 Méthodes utilisant l'apprentissage profond

Les réseaux de neurones convolutifs ou "convolutional neural networks" en anglais (couramment abrégé en CNN) font leur apparition dans les années 80 [52], [53] et [54]. Ce sont des réseaux de neurones particuliers dont le fonctionnement détaillé est expliqué dans le chapitre 3. Il s'agit d'une composition de plusieurs couches de fonctions mathématiques élémentaires. Chacune de ces fonctions prend en entrée une petite quantité d'informations et fournit en sortie des informations de plus haut niveau. Cet enchaînement de transformations permet de réaliser des tâches complexes comme de la classification d'images [55], [56] de la super-résolution d'images [57] ou encore de l'estimation de flux optique [1]. Certaines des premières couches composant les CNN sont appelées des couches de convolution. Comme leur nom l'indique, les fonctions constituant ces couches œuvrent donc à détecter la présence de motifs caractéristiques dans des signaux. Dans le cas de traitements d'images il peut par exemple s'agir de détecter les contours des objets.

Le principe de ce type réseau est d'arriver, de manière autonome, via un apprentissage, à ajuster les poids qui constituent les fonctions élémentaires pour réaliser au mieux la tâche qui lui incombe. On distingue deux grandes familles d'apprentissage : les apprentissages supervisés (par des experts) et les apprentissages auto-supervisés. C'est de cette manière que sont répartis les CNN dans la suite de cette section. Deux exceptions sont à noter : EpicFlow [58] et DeepFlow [59]. Ce sont deux réseaux ayant pour but d'estimer les flux optiques entre paires d'images suivant une architecture de type CNN mais n'étant pas soumis à une phase d'apprentissage pour ajuster au mieux les poids des fonctions dans les couches de manière autonome. En effet, les poids sont définis manuellement ce qui fait toute l'originalité de ce type de méthodes.

2.2.1 Les CNN à apprentissage supervisé

Comme son nom l'indique, un CNN à apprentissage supervisé nécessite un tuteur pour lui permettre d'apprendre et réaliser correctement la tâche pour laquelle il a été créé. Ce tuteur est, dans les faits, une base de données annotées. Chaque élément la constituant dispose de sa vérité terrain associée, objet de l'apprentissage du réseau. La base doit être suffisamment grande et variée pour que le CNN soit correctement entraîné. Parmi les plus célèbres, on retrouve par exemple la base de données MNIST (pour Modified National Institute of Standards and Technology)

[60], qui est constituée de 70 000 images de 28 × 28 pixels, représentant des chiffres écrits à la main. A chaque image est associée sa vérité terrain, c'est-à-dire, le chiffre qu'elle contient.

Hadsell et al. [61] utilisent cette base pour faire de la reconnaissance de chiffres et également pour visualiser comment un réseau s'organise et classe les données qu'on lui soumet. Plus tard, Krizhevsky et al. [55] se penchent aussi sur le fonctionnement d'un CNN. Ils remarquent qu'il fonctionne un peu comme les méthodes de détection des points d'intérêt évoquées précédemment et s'en servent pour faire de la reconnaissance d'objets. Zhao et al. cherchent d'ailleurs à intégrer la méthode SIFT à un CNN et appellent ce réseau AlphaMEX [62]. Le but de leur méthode est de proposer un système de guidage piétonnier temps réel basé uniquement sur du traitement d'image pour proposer une alternative aux systèmes GPS.

Avec un objectif plus généraliste, Dosovitskiy et al. [1] estiment le flux optique entre deux images fournies en entrée. Pour ce faire ils proposent deux réseaux dont les structures sont similaires. Ils sont composés d'une partie convolution et d'une partie déconvolution. La première partie permet de transformer et d'optimiser les informations contenues dans les images via des couches de pooling (ou couches de sous-échantillonnage). La seconde partie permet d'agrandir et de complexifier les cartes de caractéristiques obtenues à l'issue de la première partie via des couches d'unpooling (ou couches de sur-échantillonnage), donnant ainsi une estimation dense du flux optique entre les images. La particularité de ces couches d'unpooling que l'on peut aussi retrouver dans [65] est qu'elles sont composées de la concaténation de l'estimation sur-échantillonnée du flux optique de la couche précédente, de la sortie de la couche précédente et de la carte de points d'intérêt issue de la couche correspondante de la partie convolution.

La différence entre les deux CNN vient de la partie convolution. Dans la première version, appelée FlowNetSimple (abrégé en FlowNetS) les deux images sont fusionnées dès leur entrée dans le réseau. C'est ce signal qui va traverser les couches de convolution, ayant pour rôle l'extraction des caractères communs aux deux images dès le début. Dans la seconde version, appelée FlowNetCorr (pour corrélation, abrégé en FlowNetC) les deux images sont introduites séparément dans le réseau et sont traitées séparément jusqu'à ce que leurs informations soient fusionnées grâce à une couche de corrélation. Les premières couches de convolution ont donc pour but d'extraire les caractéristiques de chaque image séparément et c'est la couche de corrélation qui va essayer de trouver des correspondances entre les signaux.

Ces deux réseaux sont entraînés et testés sur une base créée pour l'occasion par les auteurs. Elle s'appelle Flying Chairs et est composée de plus de 20 000 paires d'images de taille 512 × 384 pixels. Une image est constituée d'un fond, représentant des paysages de montagnes ou de villes, issu de la base de données publique d'images : Flickr. Devant lui, sont ajoutés des modèles de chaises 3D libres de droits (809 chaises modélisées en 3D). Pour chaque image, le nombre de chaises présentes ainsi que leur disposition sont définis de manière aléatoire, donnant ainsi l'impression que les chaises volent au-dessus des paysages. Afin de générer la deuxième image de chaque paire, les chaises et le fond subissent des déplacements aléatoires et indépendants les uns des autres. Ces derniers sont toutefois soumis à quelques contraintes. Il doit s'agir de transformations affines et la moyenne leur intensité doit correspondre à

celle d'une autre base de données de déplacements artificiels : la base MPI Sintel [64]. A chaque paire vient s'ajouter la vérité terrain des déplacements d'une image à l'autre sous la forme d'une carte de flux optique, complétant ainsi la base de données Flying Chairs.

A l'issue des phases d'entraînement et de test le réseau FlowNetS semble fournir les estimations de flux optique les plus précises. Globalement, les architectures des réseaux FlowNet sont toutes deux utilisées et ont servi de base à plusieurs réseaux de la littérature. On retrouve par exemple ceux proposés par Ilg et al. [2] qui ont empilés et fusionnés les réseaux FlowNetS et FlowNetCorr pour obtenir des estimations de flux optique encore plus précises. Ils appellent d'ailleurs la meilleure combinaison de ces réseaux FlowNet 2.

Dans le domaine médical, Gaisser et al. [63] comparent les performances de leur réseau à celles proposées par le combo SIFT et RANSAC sur des images de placenta artificiel. Dans cet environnement, ils constatent que leur réseau détecte et apparie mieux les points d'intérêt pour tous les types de mouvements proposés dans la base de données qu'ils ont créé et annoté pour l'occasion.

Sur leurs bases de test. Mayer et al. [66] utilisent également les architectures FlowNet de [1] pour estimer des déplacements dans des scènes 3D. Pour ce faire ils créent une base de données appelée FlyingThings3D, composée de divers objets modélisés en 3 dimensions et évoluant dans une scène, elle aussi en 3 dimensions. Le principe reste néanmoins le même que pour les Flying Chairs, le nombre d'objets sur chaque image et leurs déplacements sont définis de manière aléatoire. Cette base de données est composée de 35 000 paires d'images, des cartes vérité terrain des flux optique et des changements de profondeur. On retrouve aussi dans cette base les cartes de profondeur de chaque image ainsi que les segmentations des volumes présents. Les réseaux qu'ils proposent sont appelés SceneFlowNet et sont des fusions des réseaux FlowNet et de réseaux DispNet qu'ils ont eux même développé pour estimer les différences de profondeurs sur les images et dont un détail de la structure est fait dans la section suivante.

L'information sur les changements de profondeur entre deux images est aussi prise en compte par Sevilla et al. [67] pour masquer les zones d'occlusion et fournir une estimation plus précise du flux optique. Ji et al. [68] se servent de l'estimation du flux optique et de la profondeur pour créer une image intermédiaire entre deux images. Enfin, Wohlhart et al. [69] estiment les changements de pose de caméra dans un environnement 3D.

Très récemment, Rau et al. [70] proposent d'estimer la profondeur d'images dans des interventions de type coloscopies. Pour ce faire, ils ont notamment recours à la création d'une base de données de vidéos artificielles d'endoscopie digestive. Ils ont en effet construit un modèle de colon en 3 dimensions et y ont fait circuler artificiellement une caméra à l'intérieur.

Parfois il n'est pas judicieux ou tout simplement impossible d'annoter une base de données suffisamment grande pour entraîner efficacement son réseau. Voilà pourquoi, un autre type d'entraînement de CNN a été développé. Leur structure et/ou leur fonction de coût permettent un apprentissage ne nécessitant pas de vérité terrain.

2.2.2 Les CNN à apprentissage auto-supervisé

Les premiers à se pencher sur l'estimation des déplacements grâce à des CNN à apprentissage auto-supervisé sont Konda et al. [71]. Pour y parvenir ils cherchent à minimiser l'erreur photométrique entre des paires d'images. La première image servant de référence et la seconde étant recalée sur le modèle de la première par le déplacement estimé. Dans l'article, ils proposent une méthode pouvant apprendre à estimer les variations de profondeurs entre deux images. Plus tard, Patraucean et al. [72] et Yu et al. [73] créent des réseaux dédiés à l'estimation dense de flux optiques entre paires d'images. Leurs architectures sont inspirées de FlowNetS . Au même moment Garg et al. [74] se penchent sur un réseau susceptible de calculer une carte de profondeur pour une image. Il s'agit encore une fois d'un réseau à apprentissage auto-supervisé mais celui-ci est un peu particulier. En effet, il prend en entrée deux images pour lesquelles la pose de la caméra est connue et également intégrée en entrée du réseau.

En 2018, Wang et al. créent un CNN à apprentissage auto-supervisé pour l'estimation bidirectionnelle de flux optique grâce à deux réseaux FlowNetS mis en parallèle [75]. En prenant toujours en entrée des paires d'images, le premier FLowNetS apprend les flux de la première image vers la deuxième et le second apprend les flux de la seconde image vers la première. Comme au paragraphe précédent, l'entraînement se fait en minimisant l'erreur photométrique entre l'image 2 recalée sur l'image 1 et l'image 1 elle même. L'estimation inverse des déplacements par le second réseau permet la mise en place d'un masque mettant en évidence les zones d'occlusions entre les deux images et rendant les cartes de flux optique proposées encore plus fidèles.

Dans leur article [76], Zhou et al. mettent aussi en place ce type de masque. Leur méthode a pour but d'estimer la profondeur de la scène dans une image et la pose de la caméra entre deux prises de vues. Le réseau qu'ils ont développé s'appelle SFM Learner. Il est lui même composé de deux réseaux principaux liés par une fonction de coût commune. Le premier CNN est de type DispNet et permet une estimation de la profondeur d'une scène à partir d'une seule image. Le second calcule la pose de la caméra à partir de paires d'images. Une branche optionnelle peut être ajoutée au second CNN permettant ainsi la création du masque évoqué précédemment.

Le réseau dédié à l'estimation de la profondeur a une structure encodeur-décodeur et prend une image en entrée. En effet, il s'agit d'estimer la profondeur d'une image uniquement à partir de celle-ci (image N). Dans l'article, on peut lire qu'une architecture complète de type FlowNetS a également été testée par les auteurs (comprenant des paires d'images en entrée) mais les estimations de profondeur réalisées n'étaient pas plus précises malgré l'ajout d'une seconde image, ce qui confirme les conclusions de Ummenhofer et al. [77].

Le réseau dédié à l'estimation de la pose de la caméra est un encodeur. Ce réseau prend en entrée des triplettes d'images, mais est structuré de la même façon que la partie encodeur de FlowNetS. Il joue d'ailleurs le même rôle, à savoir extraire des caractères communs aux images proposées en entrée. Puisque ce réseau prend trois images en entrée (N-1, N et N+1), il propose non pas une, mais deux estimation de pose, entre les images N-1 et N et entre N et N+1.

La contrainte principale de ces CNN est qu'ils sont capables de rendre des es-

timations fiables uniquement dans le cadre de mouvements rigides, c'est-à-dire que la scène est statique et que le mouvement est uniquement dû au déplacement de la caméra. Comme ce n'est pas toujours le cas, et afin de se prémunir contre des mouvements dans la scène ou de fortes variations d'éclairage, les auteurs ont mis en place une branche optionnelle dédiée à la création du masque évoqué précédemment. Ce dernier permet donc de se replacer au plus proche des conditions idéales, nécessaires au bon fonctionnement des réseaux dédiés à l'estimation de la profondeur et de la pose.

Yin et al. proposent une amélioration de [76] dans [78]. En effet, leur réseau est capable d'estimer des mouvements rigides mais également des mouvements non-rigides. C'est-à-dire des mouvements où plusieurs objets peuvent se déplacer indépendamment du mouvement principal de la caméra. On peut retrouver ce type de mouvements dans les bases de données Flying Chairs et FlyingThings3D évoquées précédemment. Leur réseau, appelé GeoNet, propose une estimation de la profondeur d'une scène, de la pose de la caméra entre deux scènes, mais aussi du flux optique entre ces deux scènes. GeoNet peut être vu comme la mise en cascade du réseau précédent et d'un réseau appelé ResFlowNet. La première partie du réseau est donc dédiée à l'estimation d'un mouvement principalement rigide. La seconde partie permet un raffinement de ce mouvement en allant détecter les zones où le mouvement n'est pas rigide. Au lieu de les masquer, elle estime les déplacements dans ces zones en calculant des flux optique de manière bidirectionnelles, comme le font Wang et al. [75] ou encore Meister et al. [79].

Enfin, Armin et al. [80] proposent un CNN pensé pour des applications médicales de types coloscopie. Il s'agit d'EndoRegNet. Ils ont choisi de s'orienter vers une méthode auto-supervisée pour palier au manque de données médicales annotées dans le secteur de l'endoscopie digestive. Leur méthode se divise en trois parties principales et donc trois sous-réseaux. Ces trois sous-réseaux partagent les mêmes entrées, à savoir des paires d'images, ainsi que la partie encodeur. Le premier sous-réseau est un décodeur visant une estimation dense du déplacement via une estimation bilinéaire du flux optique. Le second est simplement constitué de trois couches de convolution. Il a pour but de faire correspondre le déplacement estimé à une transformation polynomiale éliminant ainsi les correspondances aberrantes et permettant à la structure d'être robuste aux déformations des images. Cette branche est inspirée par [68]. Pour finir, cette architecture est, elle aussi, complétée par un masque qui est calculé par les troisième sous-réseau. Ce dernier a pour but de rendre encore plus précise l'estimation des déplacements en masquant les zones d'occlusions très présentes sur les images d'endoscopie digestive. EndoRegNet est entraîné sur plus de 45 000 paires d'images de coloscopie dont 80 pourcents sont des images issues de simulateurs. Il est testé sur plus de 1000 paires d'images, toutes issues de vraies coloscopies.

2.3 Bilan

Depuis les années cinquante de nombreuses méthodes ont été développées et améliorées. Certaines ont été utilisées dans le contexte médical et les applications principales concernent l'endoscopie digestive. Cependant, certaines études ont été

menées sur la rétine. On ne trouve en revanche pas de publication en endoscopie oculaire.

L'utilisation de méthodes classiques est développé dans le chapitre 5 de ce manuscrit tandis que le chapitre 6 présente les tests et résultats des méthodes utilisant les CNN.

A en croire l'état de l'art les méthodes basées CNN rivalisent voire surpassent les méthodes classiques. Le nombre de publications concernant l'utilisation de CNN pour l'augmentation du champ visuel à d'ailleurs beaucoup augmenté depuis 2015. Plusieurs méthodes semblent prometteuses, mais ont été publiées trop récemment pour avoir été testées dans le cadre de cette thèse.

3
Apprentissage profond

3.1 Neurones et réseaux de neurones

3.1.1 Le neurone biologique

La notion de réseaux de neurones artificiels est inspirée des réseaux de neurones biologiques. Au sein des réseaux biologiques, les neurones sont interconnectés et se transmettent des informations les uns aux autres, de proche en proche. Ils sont en quelque sorte les unités de calcul du cerveau. Chaque neurone reçoit une liste de signaux d'entrée via ses dendrites et peut générer un signal de sortie, circulant le long de son axone, après avoir effectué des calculs sur chacun des signaux d'entrée. Ces calculs permettent de déterminer l'impact des neurones les uns sur les autres. Les communications inter-neuronales se font entre l'axone du neurone précédent et une dendrite du neurone actuel via une synapse. Enfin, la sortie d'un neurone s'active si son potentiel d'action est atteint. Pour ce faire, la somme des signaux reçus en entrée (sur les dendrites) doit atteindre un certain seuil électrique.

3.1.2 Le neurone artificiel

On retrouve le même principe pour les réseaux de neurones artificiels, au sein desquels les neurones sont les unités de calcul élémentaires. Chacun d'eux reçoit en entrée un signal étant soit l'entrée du réseau (image, son ...) soit la sortie des neurones précédents.

L'impact d'une entrée peut se traduire mathématiquement par le produit de celle-ci avec un poids, puis l'ajout d'un biais. Pour un neurone ayant n entrées, la sortie sera déterminée par la somme des produits de chaque entrée par son poids associé puis l'ajout du biais. Enfin, le résultat de ce calcul est soumis à une fonction d'activation qui, comme son nom l'indique, activera ou non la sortie du neurone.

Il s'agit souvent d'une fonction non linéaire. Un schéma de neurone artificiel est proposé en figure 3.1. Dans ce schéma, w_i correspondent aux poids, x_i aux entrées, b au biais et a à la fonction d'activation. Pour une situation donnée, les paramètres définissant l'influence des neurones entre eux tels que le poids et le biais sont appris et donc évoluent au cours du temps.

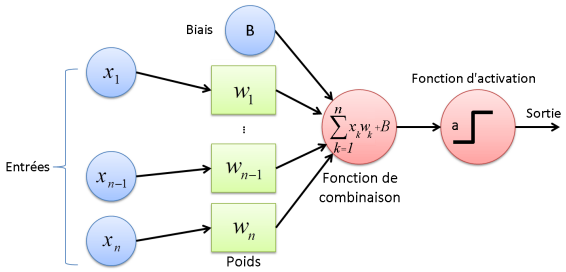


FIGURE 3.1: Schéma d'un neurone artificiel.

L'association de plusieurs neurones entre eux se fait la plupart du temps en suivant un modèle orienté acyclique. Le modèle le plus simple de réseau de neurones artificiels s'appelle le perceptron [81]. Il s'agit d'un classifieur binaire linéaire. C'est-à-dire que la sortie du réseau détermine si l'entrée proposée appartient ou non à la classe qu'il a appris à identifier. Théoriquement 1 seul neurone peut suffire pour décrire le modèle, mais dans la pratique, on ne trouve pas de modèle aussi simple. Pour séparer les deux cas, le perceptron doit passer par une étape d'apprentissage supervisé. C'est une tache au cours de laquelle le réseau apprend à estimer une fonction de prédiction à partir de données annotées (détaillé en 3.3.1). Les neurones qui ne sont pas en entrée ou en sortie du réseau sont organisé dans des couches appelées couches cachées. Un schéma de perceptron à trois entrées deux couches cachées et une sortie est proposé en figure 3.2.

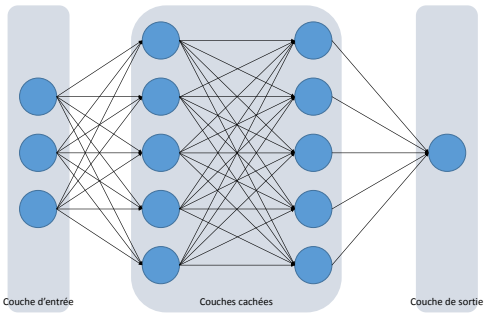


FIGURE 3.2: Schéma d'un perceptron.

Pendant l'apprentissage, la rétro-propagation est une méthode qui permet de communiquer aux neurones les erreurs qu'ils ont commises de manière à ajuster les paramètres (poids, biais) présents à chaque synapse [54]. En revanche, ce n'est pas l'erreur elle-même qui est rétro-propagée, mais son gradient. Plus l'erreur communiquée est grande, plus les poids seront modifiés de manière importante. On

retrouve essentiellement deux structures de réseaux utilisant la rétro-propagation du gradient : les perceptrons multicouches évoqués précédemment et les réseaux de neurones à convolution (ou CNN pour Convolutional Neural Networks en anglais) développés dans la partie suivante.

3.2 Le squelette des CNN

Les CNN sont des réseaux de neurones artificiels acycliques. Le motif de connexion entre les neurones est inspiré de la vision de certain animaux. En effet, dans la partie du cerveau appelée cortex visuel, leurs neurones sont disposés de manière à produire des chevauchements de l'information de leur champ visuel. Ce parallèle est approfondi dans la sous-section suivante.

La différence principale entre un perceptron multicouche vu précédemment et un CNN est la mise en commun de paramètres entre neurones permettant, via les convolutions, d'extraire des motifs caractéristiques quels que soient leur localisation dans les images (invariance par translation). Les CNN sont constitués d'un empilage de plusieurs couches (minimum 3) de neurones dont chacune a un rôle précis. Dans un CNN, on retrouve principalement 3 types de couches : les couches de convolution, les couches de mises en commun et les couches entièrement connectées.

3.2.1 La couche de convolution

La couche de convolution est à l'origine du nom de ce type de réseau et constitue l'élément de base du CNN. Elle traite, en entrée, un volume et produit, en sortie, un autre volume. On peut retrouver ce dernier sous l'appellation d'images intermédiaires ou cartes de caractéristiques. Une couche de convolution présente trois paramètres principaux : sa largeur (largeur d'une carte de caractéristiques), sa hauteur (hauteur d'une carte de caractéristiques) et sa profondeur (nombres de cartes de caractéristiques).

Chaque carte est bidimensionnelle et produite par des neurones artificiels. Leurs poids et biais peuvent évoluer au cours de l'apprentissage. Les neurones qui produisent une carte de caractéristiques commune ont tous les mêmes poids et biais. De plus, chaque neurone d'une carte couvre une partie du volume d'entrée appelée champ réceptif du neurone, comme le montre le schéma figure 3.3. C'est là que l'on retrouve la notion de chevauchement. En effet, deux neurones adjacents peuvent, en partie, couvrir la même zone d'entrée ; c'est-à-dire avoir une partie de leur champ réceptif en commun. On trouve donc l'arrivée de deux nouveaux paramètres qui sont les dimensions couvertes par un neurone (dimension du champ récepteur) et le décalage de couverture entre deux neurones (souvent appelé pas ou stride en anglais).

Dans le domaine des CNN, on remplace habituellement le mot neurone par filtre. Le fait d'attribuer à chaque filtre ayant servi à générer une carte de caractéristiques les mêmes paramètres (poids des entrées et biais) apporte d'une part de la simplicité de calcul et d'autre part de l'invariance spatiale au réseau. L'opération réalisée par un filtre est appelée convolution. Comme il s'agit d'une opération linéaire, mais que, là encore, les problèmes à résoudre peuvent être non-linéaires, la notion d'activation

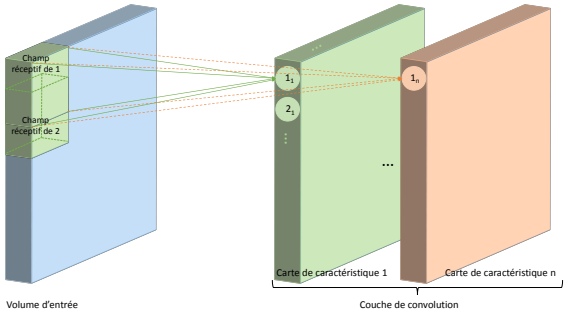


FIGURE 3.3: Schéma de principe de la couche de convolution.

est de nouveau utilisée. On parle cette fois de couche d'activation, également appelée couche ReLU (pour Rectified Linear Unit ou unité linéaire rectifiée en français). Il s'agit d'un type de fonction d'activation, mais il en existe d'autres comme la rampe, la fonction de Heaviside (ou marche d'escalier) ou la fonction Tangente Hyperbolique (voir figure 3.4).

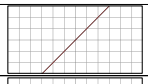
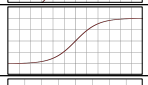
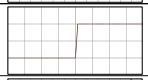
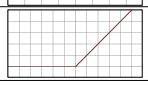
Nom de la fonction	Allure de la fonction	Equation de la fonction
Rampe		$f(x) = x$
Tangente Hyperbolique		$f(x) = \tanh(x)$
Heaviside		$f(x) = \begin{cases} 0 & \text{pour } x < 0 \\ 1 & \text{pour } x \geq 0 \end{cases}$
ReLU		$f(x) = \begin{cases} 0 & \text{pour } x < 0 \\ x & \text{pour } x \geq 0 \end{cases}$

FIGURE 3.4: Quelques fonctions d'activation usuelles.

3.2.2 La couche de mise en commun

On la retrouve aussi sous le nom de couche de pooling. Elle a pour but de sous-échantillonner le signal. Dans le cas d'une image, celle-ci est découpée en fenêtres ayant toutes les mêmes dimensions et ne se chevauchant pas. Le sous-échantillonnage se fait en ne retenant qu'une valeur par fenêtre. La plupart du temps c'est la valeur maximale de chaque fenêtre qui est retenue (voir figure 3.5). C'est ce qu'on appelle une couche de max pooling. On peut aussi choisir de retenir la valeur moyenne (average pooling) ou la norme L2 (L2-norm pooling). Cette couche modifie donc la hauteur et la longueur du signal, mais pas sa profondeur. Elle est utilisée pour rendre l'apprentissage plus rapide et permet l'extraction de caractéristiques à différentes résolutions spatiales. Enfin, elle réduit le nombre de paramètres du réseau limitant ainsi le sur-apprentissage. Habituellement, ce genre de couche se trouve intercalée entre deux couches de convolutions.

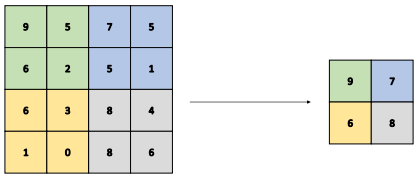


FIGURE 3.5: Schéma de principe de la couche de mise en commun.

3.2.3 La couche entièrement connectée

La couche entièrement connectée ou couche dense est plus communément retrouvée sous son appellation anglaise de couche "fully connected". Tous les neurones de cette couche ont leurs entrées connectées aux sorties de la couche précédente et leur sortie connectées aux entrées de la couche suivante. Le but de cette couche est de réaliser des tâches de classification. On les retrouve souvent en fin de réseau. Dans un contexte de classification, si cette couche est placée en dernier, c'est elle qui produit les probabilités d'appartenance aux différentes classes.

Les combinaisons de ces trois types de couches principales permettent de créer des réseaux. Même si l'agencement des couches est important dans l'efficacité d'un réseau, celui-ci ne se révèle utile qu'une fois entraîné. Il doit donc passer par une étape d'apprentissage qui peut être de plusieurs types.

3.3 Les principaux types d'apprentissage

Il existe deux manières principales de concevoir l'entraînement d'un réseau. La première partie de cette section détaille la méthode dans laquelle l'apprentissage de fait de manière guidée. La seconde partie traite d'une technique où le réseau s'entraîne de manière plus autonome. Enfin, la troisième partie porte sur l'intérêt de réaliser un apprentissage en deux temps.

3.3.1 L'apprentissage fortement supervisé

Dans le domaine de l'apprentissage automatique, l'apprentissage fortement supervisé consiste à faire apprendre à un réseau, une fonction de prédiction (ou modèle) grâce à une base de données de cas annotés. Les cas de la base doivent être suffisamment nombreux, variés et représentatifs pour que le modèle appris soit efficace. Pendant l'apprentissage, les différents exemples contenus dans la base sont fournis en entrée du réseau et les annotations qui servent de vérité terrain sont fournies en sortie du réseau. C'est à ces annotations que le réseau va comparer la sortie qu'il a estimée. L'objectif final étant, pour chaque cas, de minimiser l'écart entre la sortie du réseau et la vérité terrain associée. Les paramètres du réseau sont itérativement mis à jour de manière à ce que la fonction de coût soit minimisée. Une formule générale

de mise à jour des poids est proposée dans l'équation (3.1) où w_i sont les poids, α le taux d'apprentissage, et *loss* la fonction de coût. On parle alors de minimisation de fonction de coût. Ainsi, à l'issue de l'entraînement, le réseau doit être capable de rendre un verdict correct pour des cas non présentés pendant l'apprentissage et dont les vérités terrains ne lui sont pas fournies.

$$w_i \leftarrow w_i \ - \ \alpha \ \frac{d \ loss}{d \ w_i} \tag{3.1}$$

Le verdict proposé peut se présenter sous deux formes principales. La première est la classification. La sortie se présente sous la forme d'une valeur discrète, appartenance ou non à une catégorie dans le cas d'une classification binaire ou appartenance à telle ou telle catégorie dans le cas d'une classification multi-classes. La seconde est la régression. La sortie se présente sous la forme d'une valeur continue, c'est-à-dire qu'on cherche à faire correspondre à chaque entrée un ou plusieurs nombre (comme par exemple un coût, une probabilité, une variation...). On retrouve également ces deux types de prédiction pour les méthodes utilisant des apprentissages auto-supervisé.

Parfois, il n'est pas judicieux ou tout simplement impossible d'annoter une base de données suffisamment grande pour entraîner efficacement son réseau. Voilà pourquoi, un autre type d'entraînement de CNN a été développé. Leur structure et/ou leur fonction de coût permettent un apprentissage ne nécessitant pas de vérité terrain.

3.3.2 L'apprentissage auto-supervisé

En apprentissage auto-supervisé, les bases de données ne sont donc pas annotées. Le réseau se charge lui-même de trouver des différences ou des similarités entre les données. C'est lui qui va fixer les catégories dans le cas de classification ou estimer des valeurs dans le cas de régression.

L'optimisation des paramètres du réseau se fait toujours grâce à la minimisation d'une fonction de coût. Cependant, comme il n'y a pas de vérité terrain à laquelle se comparer, la plupart du temps, dans la fonction de coût, la sortie du réseau est comparée à l'entrée de l'itération actuelle, suivante ou précédente.

C'est-à-dire que ce type de réseau est capable de réaliser une tâche sans jamais qu'on lui ait montré le/les résultat/s à produire. L'exemple le plus connu de ce type d'application concerne les langues et plus précisément la mise en place d'un traducteur par Facebook [82]. Celui-ci permet de traduire un texte d'une langue à une autre, et même de prédire les mots d'une conversation, sans jamais avoir vu d'exemple concret de mots traduits d'une langue à une autre. Pour ce faire il va mettre analyser les contextes dans lesquels sont utilisés les mots dans différentes langues et ainsi établir des correspondances.

3.3.3 L'apprentissage par transfert

L'apprentissage par transfert (ou transfer learning en anglais) est une technique d'apprentissage automatique qui consiste à utiliser des connaissances préalablement

acquises, par le même réseau, grâce à un apprentissage précédent sur une nouvelle étude [83]. Il est envisageable en apprentissage fortement supervisé aussi bien qu'en apprentissage auto-supervisé. En apprentissage fortement supervisé, le recours à cette méthode ce fait généralement lorsque peu de données annotées sont disponibles pour l'étude à réaliser tandis qu'un grand nombre est disponible pour une autre étude.

L'utilisation de l'apprentissage par transfert est donc une tentative de démarrage du processus d'apprentissage d'un modèle à partir d'un modèle appris pour une tâche différente. Dans le cas d'un CNN, au lieu de commencer l'apprentissage de zéro, il y a donc une initialisation des paramètres du réseau qui permet une convergence plus rapide du modèle ciblé. En pratique, entraîner un CNN de zéro sur une base de données complexe est relativement rare en raison de la complexité de la tâche à réaliser et du nombre limité de ressources disponibles. En revanche, il est possible et plus aisé de peaufiner l'état d'un CNN, pré-entraîné sur un très grand jeu de données initial, sur le jeu de données cible.

Cette étape d'initialisation sert à faire apprendre au réseau des caractéristiques génériques qu'il faut retrouver dans toutes les situations comme par exemple la détection de contours. Il s'agit plutôt d'apprentissage des premières couches du réseau. En effet, dans un CNN, plus une couche est profonde, plus son rôle devient spécifique à une certaine tâche ou type d'images.

Deux approches d'apprentissage par transfert sont envisageables. La première consiste à pré-entraîner l'intégralité du réseau sur la première base de données dans un premier temps. Puis, dans un second temps, raffiner les couches denses du réseau (les dernières couches) avec la seconde base de données et figer les poids et biais des premières couches. Cette approche concerne essentiellement les réseaux de type classifieur pour lesquels il y a peu de données annotées. Se faisant, la modification sur les couches denses va permettre au réseau de se spécifier sur l'opération de classification désirée, sans pour autant perdre l'apprentissage des premières couches de convolution plus génériques. La deuxième approche consiste à inclure tout, ou une partie des couches de convolution dans le raffinement. Cette approche plus naïve et généraliste n'est efficace que si la seconde base de données est assez grande pour modifier efficacement l'apprentissage du réseau.

En réalité il existe d'autres types d'apprentissage, ils sont historiquement moins utilisés, mais depuis quelques années leur utilisation devient plus courante. Il s'agit de l'apprentissage semi-supervisé et de l'apprentissage par renforcement. Dans le premier cas, l'apprentissage se fait à partir de données annotées et non annotées, d'où son nom. Dans le second cas, l'apprentissage se base sur la répétition d'une expérience, dans un certain environnement, soldée par une récompense ou un échec. Le but de cette méthode étant évidemment de maximiser la récompense et de minimiser l'échec. La récompense maximale correspond au succès de l'expérience dans son intégralité. L'exemple le plus commun est celui d'une intelligence artificielle qui essaie de terminer par elle-même le niveau d'un jeu vidéo. Dans ce cas, l'échec se traduit par la "mort du personnage" ou la fin du timer et le succès par l'accomplissement du niveau.

3.4 Différents types de CNN

Dans cette thèse, deux types de CNN ont été étudiés, les encodeurs-décodeurs et les auto-encodeurs cependant il existe beaucoup d'autres architectures de réseaux qui sont détaillées sur ce site (*https ://www.asimovinstitute.org/author/fjodorvanveen/*). La section suivante détaille le fonctionnement des encodeurs, celui des décodeurs puis des deux types de CNN utilisant ces architectures.

3.4.1 Les encodeurs

Les encodeurs sont une catégorie de réseau de neurones dont l'objectif est d'apprendre (de manière supervisée ou non) une représentation spécifique à partir d'une base de données d'images. Cette représentation est obtenue en sortie du réseau sous la forme d'un vecteur/carte de caractéristiques tandis qu'il prend en entrée une ou plusieurs images. Dans le cas d'un entraînement réussi, elle contient les caractéristiques nécessaires et suffisantes pour décrire et identifier les données d'entrée.

La structure d'un CNN encodeur est typiquement composée de la répétition de plusieurs couches de convolution/activation et d'une couche de pooling. Enfin, on trouve les couches denses (généralement entre une et deux). Le schéma en figure 3.6 représente un exemple d'architecture d'encodeur.

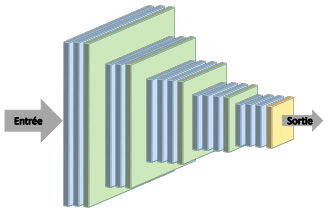


FIGURE 3.6: Schéma d'un CNN encodeur.
En bleu : Couche de convolution/activation - En vert : Couche de pooling - En jaune : Couche dense

3.4.2 Les décodeurs

A l'inverse, les décodeurs prennent en entrée un vecteur/carte de caractéristiques et proposent en sortie un signal décodé du même type que le signal à l'entrée de l'encodeur (une image). Contrairement aux réseaux encodeurs que l'on peut exploiter seuls, les décodeurs sont toujours précédés d'un encodeur.

Généralement un CNN décodeur commence par une couche appelée couche d'unpooling (ou de sur-échantillonnage). Elle ressemble à la couche de pooling, mais réalise l'opération inverse. En effet, elle a pour but de modifier la hauteur et la largeur du signal sans modifier sa profondeur. Pour ce faire, elle découpe son entrée en fenêtres élémentaires ayant toutes les mêmes dimensions et décompose chaque

fenêtre en n par n nouvelles fenêtres (Comme le montre la figure 3.7). Cette couche précède un enchaînement de plusieurs couches de convolution/activation. Comme pour l'encodeur, le motif couche d'unpooling puis couches de convolution/activation peut se répéter plusieurs fois. Il n'est pas rare de voir un décodeur organisé de manière symétrique à l'encodeur qui le précède. En fonction du type d'apprentissage utilisé, l'association de ces deux structures peut engendrer deux types de réseaux. Les encodeurs-décodeurs et les auto-encodeurs.

3.4.3 Les encodeurs-décodeurs

Comme leur nom l'indique, les encodeurs-décodeurs sont composés d'une première partie encodeur et d'une seconde partie décodeur. Ce type d'architecture s'entraîne généralement grâce à un apprentissage supervisé, donc avec des bases de données annotées. Un exemple d'encodeur-décodeur est proposé en figure 3.7.

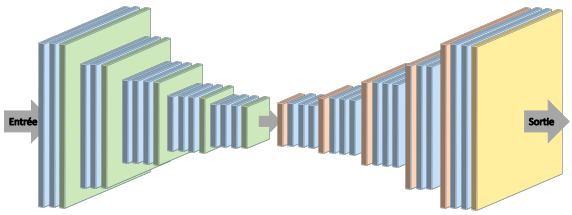


FIGURE 3.7: Schéma d'un CNN encodeur-décodeur.
En bleu : Couche de convolution/activation - En vert : Couche de pooling - En orange : Couche d'unpooling - En jaune : Couche dense

3.4.4 Les auto-encodeurs

Les réseaux de neurones auto-encodeurs appartiennent à la catégorie des réseaux à apprentissage auto-supervisés. En effet, il n'y a pas besoin de fournir de données annotées pour entraîner efficacement ce type de réseau. La partie encodeur se charge de transformer les données en vecteurs de caractéristiques et la partie décodeur se charge d'essayer de reconstruire le signal d'entrée à partir des vecteurs de caractéristiques. Ainsi, l'entraînement du réseau se fait en minimisant l'écart entre le signal à l'entrée de l'encodeur et celui estimé en sortie du décodeur. La plupart du temps, seule la sortie de l'encodeur est utilisée par la suite et la sortie du décodeur sert simplement à l'entraînement du réseau.

De ce type d'architecture découle plusieurs variantes comme les auto-encodeurs débruiteurs, épars, variationnels ou encore contractifs. Ces variantes ne sont pas développées dans ce manuscrit. Pour plus d'information sur ces structures, se référer à l'article de Bengio et al. [84].

3.5 Bilan

Au fil du temps les réseaux de neurones artificiels se sont diversifiés et complexifiés. Avec la mise en commun de paramètres entre neurones d'une même couche grâce à des convolutions, on note l'apparition des CNN. Leur utilisation a également évolué et on les retrouve dans d'autres tâches que de la classification. Enfin, on remarque que les modes d'apprentissages se sont aussi diversifiés.

En revanche, il est intéressant de noter que peu importe le style d'apprentissage, le rôle de la base de données utilisée est primordial. Des facteurs comme la qualité, la quantité ou encore la diversité des données qui la composent constituent des clés pour l'apprentissage d'un réseau.

4

Acquisition de données

L'objectif d'un traiteur d'images est d'apporter de la valeur ajoutée aux données qui lui sont soumises. Il peut le faire de plusieurs manières. D'une part, il peut tenter d'en extraire de l'information (présence, caractéristiques d'un objet) pour faciliter une décision prise par la suite. D'autre part, il peut tenter de modifier l'image pour augmenter la qualité des informations qu'elle contient. Il ne faut pas pour autant négliger l'importance de la phase d'acquisition des images. Ainsi, des leviers comme les conditions d'acquisition ou le matériel utilisé sont primordiaux dans la chaîne de traitement d'informations.

L'endoscope oculaire et la lampe à fente sont deux dispositifs de visualisation utilisés notamment pour faciliter les interventions chirurgicales de la chambre postérieure de l'œil, qui est le segment étudié dans le cadre de cette thèse. Cette zone commence après le cristallin et va jusqu'à la rétine (Schéma Figure 4.1).

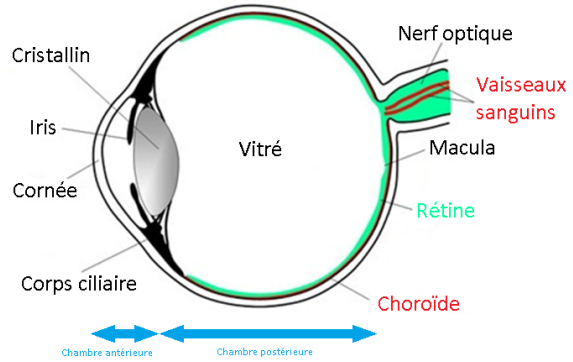


FIGURE 4.1: Schéma d'un œil.

Les deux premières sous-sections de ce chapitre sont organisées de la même

manière : développement d'un historique de l'outil puis détails de la base de données issue de cet outil. La première porte sur la lampe à fente et la seconde sur l'endoscope oculaire. Enfin, une troisième sous-section aborde les autres bases de données créées et/ou utilisées dans la suite du manuscrit.

4.1 La lampe à fente

4.1.1 Historique

La lampe à fente, que l'on peut également trouver sous le nom de biomicroscope est un dispositif qui a été mis en place pour la première fois en 1911 [85] par l'ophtalmologue suédois Allvar Gullstrand également connu pour sa formule sur la vergence d'un système optique et pour avoir obtenu un prix Nobel en 1911 pour ses travaux sur les dioptries de l'œil.

Cet outil permet, comme précisé précédemment, de visualiser la chambre postérieure de l'œil, mais aussi la chambre antérieure et également sa surface. En effet, son utilisation primaire était à l'origine principalement faite autour de la chambre antérieure, mais l'arrivée de certaines lentilles dans les années 80 [86] rendent son utilisation courante pour des examens de la rétine. Au fil du temps, ce dispositif s'est vu enrichi de plusieurs outils lui permettant de réaliser des mesures comme la profondeur de l'œil avec un laser [87]. Elle peut aussi être couplée à un OCT (Optical Coherence Tomography) [88], [89] pour visualiser en profondeur les éléments de la chambre antérieure. Mesurer la pression intraoculaire est une manipulation importante pour le dépistage de certaines pathologie voilà pourquoi on retrouve des lampes à fente équipées de tonomètres [90]. La même année, un système échographique lui a même été ajouté dans [91] pour avoir des informations sur la chambre postérieure dans le cas de pathologies rendant le cristallin trop opaque pour voir à travers. Cette grande diversité de visualisations et de mesures en fait un outil indispensable pour les ophtalmologistes et permet une aide au diagnostic pour de nombreuses pathologies oculaires.

Fondamentalement, un biomicroscope est composé de trois éléments principaux [92]. On trouve en premier lieu la source de lumière, appelée lampe à fente et d'où tire son nom l'objet. Elle produit un faisceau de lumière dont la hauteur, la largeur ainsi que la position peuvent être modifiés. La plupart des lampes à fente actuelles sont construites à partir de diodes électro luminescentes (ou LED : Light-Emitting Diode) dont la longueur d'onde peut être variable ou modifiée en aval grâce à des filtres. Elles sont conçues et disposées de sorte à éclairer de manière optimale le second élément.

Il s'agit du microscope binoculaire. Son grossissement doit bien entendu être variable pour s'adapter aux types d'examens à réaliser. Les grossissements les plus courants sont $\times 10$, $\times 16$ et $\times 25$. De plus, pour un meilleur confort d'utilisation, le champ visuel ainsi que la profondeur de champ doivent être réglables et aussi grands que possibles. La distance entre les optiques de l'objectif et l'œil du patient doit être assez grande pour que celui-ci puisse, si nécessaire, être manipulé par l'ophtalmologiste.

Enfin, un dispositif mécanique sert à la fois de liaison entre la lampe et le micro-

scope et de station d'accueil pour le patient. Il est composé d'un arbre qui permet de pouvoir bouger de manière indépendante le microscope et la lampe. Cet arbre est lié à un plateau mobile qui peut se déplacer horizontalement et verticalement pour que l'œil du patient puisse être positionné correctement en face de l'objectif. Sur les lampes à fente récentes ces paramètres sont réglables électroniquement via des manettes.

Au fil du temps, ce dispositif s'est adapté à son époque en intégrant de plus en plus d'électronique, en réalisant des acquisitions de manière numérique ou encore en intégrant des programmes pour rendre la procédure plus confortable. Ainsi, plusieurs équipes de recherche ont tenté de créer des mosaïques d'images de rétines acquises par lampe à fente, en 2D [93], [18] ou en 3D [94]. D'autres travaux ont pour objectif l'aide au diagnostic comme la détection et la classification de cataractes [95] ou [96]. Enfin, certaines publications explorent des pistes différentes comme [97] qui compare des clichés acquis par lampe à fente classique à des clichés acquis via un smartphone équipé d'un objectif particulier appelé D-eye. La comparaison a pour but d'étudier la viabilité d'une campagne de dépistage massive et bas coût de la rétinopathie diabétique.

4.1.2 Les données issues de la lampe à fente

L'entreprise Quantel nous a fourni trois vidéos acquises à la lampe à fente dont un modèle est proposé figure 4.2. La première dure 5 minutes et 4 secondes, la seconde ne dure que 26 secondes et la troisième dure 1 minute et 51 secondes (voir bas du tableau 4.1). Les trois vidéos ont une définition de 1 280 × 720 pixels et une fréquence d'acquisition de 60 images par seconde. Nous savons également que les vidéos ont été acquises en 2014. En revanche, l'entreprise ne nous a pas communiqué d'information sur le modèle de lampe à fente utilisé ou sur les conditions d'acquisition.

FIGURE 4.2: Lampe à fente.
Image extraite du site : http ://www.medicalexpo.fr

Dans ces vidéos, la fente balaye la rétine, mais l'intégralité de chaque vidéo n'est pas exploitable (un récapitulatif des données exploitables détaillé dans ce paragraphe est proposé dans la partie inférieure du tableau reftab :recapdata). En effet, dans la première vidéo, le patient ferme l'œil à quatre reprises pour des durées allant de 4 à 17 secondes. De plus, il semble y avoir des problèmes à l'enregistrement de la vidéo, car par deux fois, l'image se gèle quelques secondes sur une zone de la rétine et la vidéo reprend sur une autre zone. Enfin, dans les dernières quinze secondes de la vidéo, l'enregistrement continue alors que l'acquisition est terminée, donnant une séquence noire. Ces trois phénomènes se retrouvent également dans la troisième vidéo et dans des durées similaires.

Comme nous cherchons à estimer les déplacements entre deux images de rétines, nous devons traiter des images exploitables présentant des rétines. C'est donc naturellement que nous avons divisé la base en deux catégories : images exploitables et images non-exploitables. Par la suite, nous n'utiliserons que la première catégorie. La partie détection automatique de l'exploitabilité d'une image pourra faire l'objet de recherches futures. En ne considérant que les images exploitables, nous arrivons à une base de données de plus de 14 0000 images (14 215).

Il peut également être utile de noter que l'intégralité de l'image n'est pas exploitable. Comme précisé précédemment, la fente présente la particularité de pouvoir changer de taille et se déplacer légèrement ce qui aboutit à la présence de bandes noires sur l'image et au déplacement de la zone d'intérêt, comme le montre les images figure 4.3. La largeur moyenne de la fente est de 394 pixels et on considère qu'elle occupe en permanence la totalité de la longueur des images (soit 1 280 pixels).

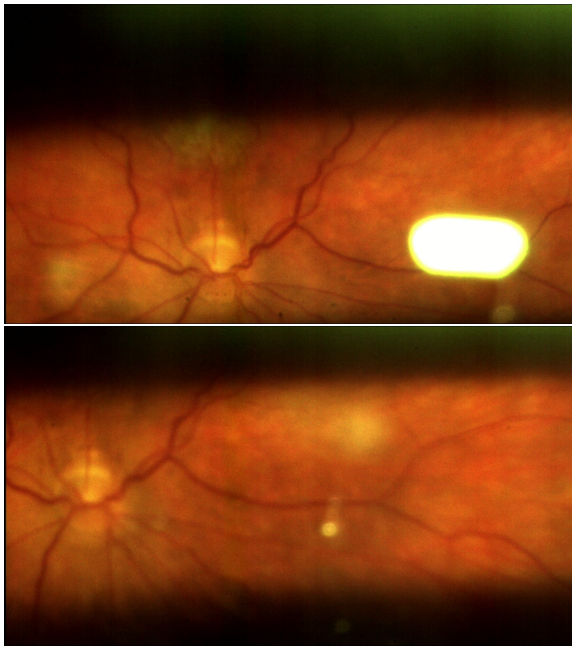


FIGURE 4.3: Deux images acquises par la lampe à fente (fournies par Quantel) illustrant le déplacement de la fente.

4.2 L'endoscope oculaire

4.2.1 Historique

L'endoscopie est une technique d'imagerie médicale mini-invasive qui permet de visualiser l'intérieur du corps. Le premier concept d'endoscope oculaire est proposé par Thorpe en 1934 dans [98]. Il propose un tel outil pour pouvoir retirer de l'œil des corps étrangers s'y étant incrustés et dont le seul moyen de les localiser est d'avoir une vision interne de l'organe. Son prototype est composé d'un télescope galiléen qui joue le rôle d'objectif, d'une lampe qui sert de source de lumière, de forceps pour retirer les corps étrangers et d'un oculaire amovible. L'objectif, les forceps et la source de lumière étant destinées à rentrer dans l'œil, il parvient à les contenir dans une gaine rigide de 6.5 mm de diamètre (soit 2 Gauge). Une fois la gaine insérée via une incision de 8 mm dans la sclère, il est possible de fixer l'oculaire à cette gaine ainsi que le raccord électrique pour la lampe.

En 1952, Butterworth et al. proposent également un endoscope oculaire [99], où l'on retrouve l'association objectif, oculaire et source de lumière. En revanche, sur leur modèle, la gaine ne fait plus que 2.5 mm de diamètre (10 Gauge) et les forceps disparaissent, permettant une introduction plus aisée et des incisions moins larges. Au fil du temps, les matériaux évoluent et l'outil aussi. Ainsi, en 1978, Norris et al. [100] proposent un endoscope dont la source de lumière passe par des fibres optiques amenant l'outil à un diamètre de 1.7 mm (environ 14 Gauge). A cette époque, l'outil est encore entièrement rigide.

Il faut attendre 1990 et les travaux de Volkov et al. [101] pour voir apparaître les premiers modèles d'endoscopes oculaires flexibles. La gaine qui rentre dans l'œil reste toujours rigide et mesure 1.2 mm de diamètre (environ 16 Gauge). En revanche, la partie entre l'objectif et l'oculaire devient flexible rendant la procédure moins contraignante pour le chirurgien. Toujours dans l'objectif de rendre l'intervention plus facile pour le chirurgien et moins invasive pour le patient, Eguchi et al. [102] conçoivent un outil dont le diamètre est de 0.8 mm (soit 20 Gauge) et remplacent l'oculaire par un capteur CCD (Charges-Coupled Device). Cet ajout permet une visualisation sur moniteur, donnant la possibilité à plusieurs personnes de suivre aisément l'intervention en temps réels.

Enfin, en 1992, Uram [103] introduit un laser dans la gaine de l'outil apportant à l'endoscope la fonction de photo-coagulation, utile pour certaines chirurgies. L'intégralité des signaux (éclairage, laser et images) qui transitent dans la gaine passe par des fibres optiques. Son diamètre est toujours de 0.8 mm, le capteur CCD remplace toujours l'oculaire. De plus, le champ visuel proposé est de 70 degrés.

De nos jours, le principe reste inchangé et les endoscopes utilisés sont très similaires à celui de [103]. Des améliorations sont toutefois à noter en terme de diminution de consommation d'énergie, d'augmentation de puissance du laser, d'augmentation du champ visuel proposé et de qualité des capteurs/moniteurs [104]. Des travaux ont aussi été menés pour diminuer d'avantage le diamètre de l'outil, au détriment de la largeur du champ visuel et de la qualité de l'éclairage [105].

Les recours à cet outil se font principalement lorsque la chirurgie traditionnelle sous microscope n'est pas possible (trop forte opacité d'un ou plusieurs éléments du segment antérieur ou zone à opérer en dehors du champ visuel accessible grâce au

microscope). Dans ces situations, l'endoscope oculaire peut directement être inséré dans, ou à proximité de, la zone à traiter donnant ainsi un visuel au chirurgien.

Une procédure courante se faisant sous endoscope oculaire est la photo-coagulation du corps ciliaire pour traiter le glaucome [106], [107]. L'endoscope permet une bonne visualisation de cette zone par rapport au microscope. De plus, le laser présent sur l'endoscope permet de détruire une partie du corps ciliaire qui produit l'humeur aqueuse. Si cette dernière est produite en trop grande quantité ou est mal évacuée de l'œil elle va faire augmenter la pression intraoculaire et provoquer des dommages sur le nerf optique (donc un glaucome). Bien que plus rares, on peut aussi trouver des chirurgies de la cataractes faites sous endoscopie oculaire [108], [109].

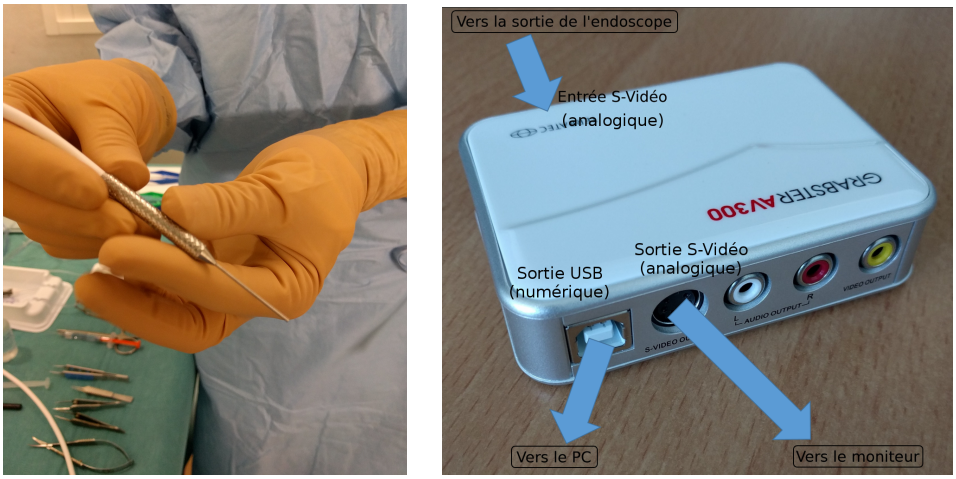
Les autres procédures concernent la chambre postérieure. Les chirurgies sous endoscope oculaire peuvent se faire pour traiter des proliférations vitréo-rétiniennes (PVR) [110] faisant suite à un décollement de rétine [111]. On trouve aussi plusieurs cas de traumas qui ont été pris en charge par endoscopie oculaire [112]. Enfin, l'endoscopie oculaire est également utilisée pour l'extraction de corps étranger [113] ayant causé des endophtalmies [114], pathologie à l'origine de la création de cet outil de visualisation.

4.2.2 Les données issues de l'endoscope oculaire

Dans cette thèse, les vidéos exploitées ont été acquises avec un endoscope oculaire Endo Optiks Ome 200 droit (Figure 4.4). Il possède un diamètre de 0.8 mm (20 Gauge) et est composé de fibre optique. Celles-ci servent à transmettre la lumière de la source lumineuse, à capter les images et également à faire circuler un flux laser. Au bloc opératoire, le système de visualisation se fait autour d'un moniteur analogique et toutes les sorties du boîtier de l'endoscope sont analogiques. Pour enregistrer des vidéos numériques, il a donc fallu convertir une partie du signal de sortie. Pour ce faire, nous avons eu recours à deux appareils différents.

Le premier est un convertisseur analogique numérique de la marque Terratec. Il s'agit du modèle Grabster AV300 MX. Ce module permet de rentrer un signal analogique via un câble S-Vidéo et de sortir à la fois un signal analogique via une sortie S-Vidéo et un signal numérique via une sortie USB (Figure 4.4). Le logiciel d'acquisition est le Video Easy de la marque Magix. Il permet des enregistrements de vidéos de résolution 720×576 pixels. Trois vidéos ont été acquises avec ce dispositif. La première traite un glaucome et les deux autres des endophtalmies. Elles ont des durées respectives de 18 minutes et 2 secondes, 7 minutes et 43 secondes, et 3 minutes et 12 secondes. Leur fréquence d'acquisition est identique et égale à 25 images par seconde.

Après plusieurs tentatives de réglages, nous avons conclu que ce boîtier permettait de faire de bonnes acquisitions de chirurgies du glaucome, mais pas des autres chirurgies de la chambre postérieure (Figure 4.5). En effet, malgré toutes nos tentatives de réglages de la luminosité per et post enregistrement, les vidéos sont trop sombres pour être exploitables comme le montrent les figures b et c de 4.5. En effet, nous avons constaté que l'utilisation de ce dispositif détournait une partie du signal destiné à la visualisation sur le moniteur. En augmentant la luminosité sur les enregistrements nous diminuions celle sur le moniteur, rendant la chirurgie



(a) (b)

FIGURE 4.4: Outils utilisés pour l'acquisition des vidéos. a. Endoscope oculaire Endo Optiks Ome 200 - b. Exemple de convertisseur analogique numérique

(a) (b) (c)

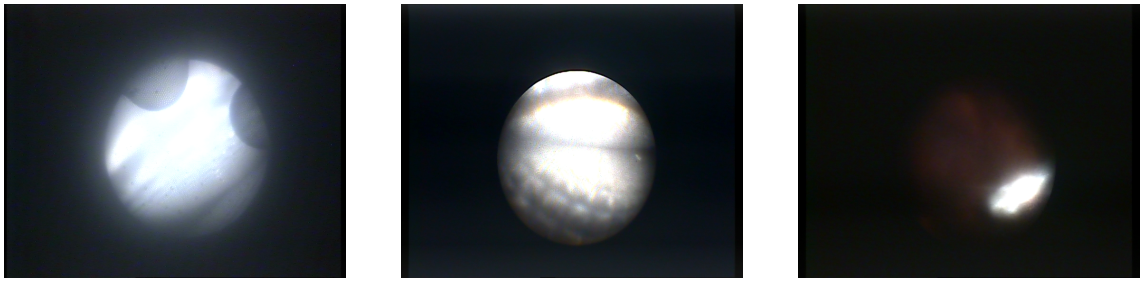
FIGURE 4.5: Extraits des 3 premières acquisitions. a. Première vidéo - b. Deuxième vidéo - c. Troisième vidéo

plus complexe pour le chirurgien. Nous n'avons donc naturellement pas retenu cette solution d'acquisition.

Le second boîtier est aussi un convertisseur analogique numérique. Il s'agit du modèle Intensity Shuttle USB de la marque Blackmagic Design. Il s'agit d'un modèle supérieur en gamme au Grabster AV300 MX. Le logiciel d'acquisition s'appelle Media Express et est lui aussi de la marque Blackmagic. L'utilisation de ce système d'acquisition donne également des vidéos de résolution 720×576 pixels.

Avec le second boîtier, 25 vidéos ont été enregistrées. Cette fois, le dispositif assure une image plus exploitable, sans assombrir le signal du moniteur. Comme pour les vidéos acquises à la lampe à fente, nous avons fait un tri manuel dans les images afin de déterminer quelles images étaient exploitables et lesquelles ne l'étaient pas. Là encore, effectuer un tel tri de manière automatique pourra faire l'objet de travaux futurs. Un récapitulatif autour de ces vidéos est proposé dans la suite de cette section ainsi que dans le tableau 4.1.

Sur les 25 vidéos, deux n'ont pas pu être utilisées pour des problèmes extrêmes de réglages de luminosité ou parce que l'instrument n'est quasiment pas utilisé, c'est-à-dire que l'endoscope reste allumé, posé sur la table, mais rentre très brièvement



(a) (b) (c)

FIGURE 4.6: Images d'endoscope à exclure de la base. a. Outil à l'extérieur de l'œil (1) - b. Outil à l'extérieur de l'œil (2) - c. Éclairage pas encore réglé

dans l'œil (Figure 4.6). La base de données est donc constituée à partir de 23 vidéos. Là encore, on retrouve une chirurgie du glaucome. Elle fait partie des plus courtes et dure 7 minutes. On compte un traitement d'hémorragie par photo-coagulation. Celui-ci débute par une étape de vitrectomie. La vidéo dure 32 minutes. Cependant, à partir de 20 minutes, l'endoscope sort de l'œil, n'y rentre plus mais l'enregistrement continue.

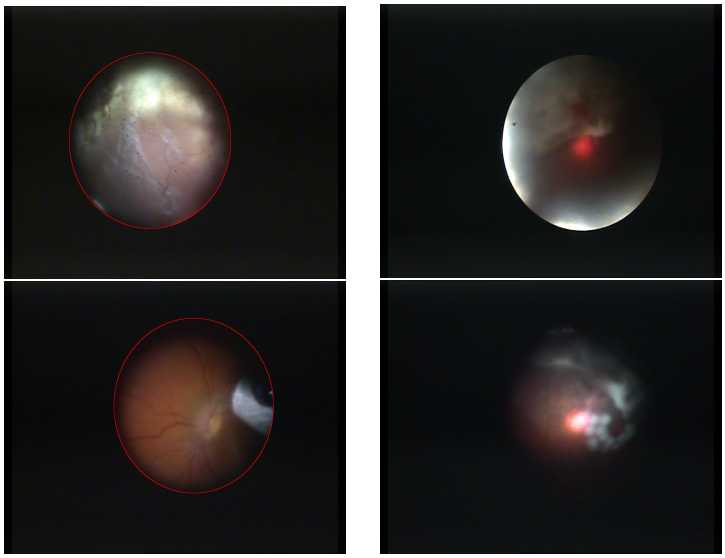
On trouve également deux vidéos de traitement de traumatismes, l'une de 51 minutes et l'autre de 21 minutes. Encore une fois, on trouve plusieurs minutes pendant lesquelles l'enregistrement se fait sans que l'endoscope soit dans l'œil. Cette situation peut se produire au début, pendant ou à la fin des enregistrements pour des durées allant de 2 à plus de 10 minutes.

Parmi les 23 vidéos, deux traitent des endophtalmies et montrent des vitrectomies. Elles durent respectivement 26 et 35 minutes. Dans ces enregistrements, l'endoscope est présent dans l'œil sur la totalité des images.

Les 17 autres enregistrements traitent des décollements de rétine. Le plus court dure 8 minutes et le plus long 64. Tous contiennent un passage de vitrectomie, 6 ont une phase de pelage de membrane, 3 ont une phase de cerclage et 4 ont une phase de photo-coagulation. Dix vidéos débutent et/ou se terminent par des passages où l'endoscope n'est pas dans l'œil, laissant penser que la chirurgie est déjà commencée ou pas encore terminée. Ces passages peuvent durer jusqu'à 12 minutes. Enfin, dans 9 vidéos, il arrive que l'outil sorte de l'œil pendant la chirurgie avant d'y ré-entrer sans pour autant que l'enregistrement soit coupé. Ces sorties varient entre 1 et 9 selon les vidéos et dure d'une dizaine de secondes à 7 minutes.

Comme nous l'avons vu, la durée des vidéos est donc très variable (entre 7 et 64 minutes). Les plus courtes correspondent à des chirurgies ou le recours à l'endoscopie n'est que ponctuel ou que l'opération réalisée reste basique comme par exemple la réalisation d'une légère vitrectomie. A l'inverse les vidéos les plus longues correspondent à des chirurgies entières et plus complexes réalisées par endoscopie. Dans celles-ci, on retrouve toujours une étape de vitrectomie en début de chirurgie, mais elle est suivie d'autres étapes plus longues comme un pelage de membrane, une photo-coagulation au laser ou encore un cerclage.

La durée moyenne des vidéos est de 35 minutes. Sur les 750×576 pixels de l'image, plus des deux tiers n'est pas utilisée. En effet, les images sont de formes rectangulaires, mais le signal proposé par l'endoscope est plutôt assimilable à une



(a) (b)

FIGURE 4.7: Images issues de l'endoscope oculaire. a. Deux exemples de zones utiles - b. Déplacement de la zone utile pendant une même vidéo

ellipse. Celle-ci est aussi appelée zone utile dans la suite du manuscrit. Les fibres optiques étant souples et légèrement mobiles à l'intérieur de la gaine, des manipulations comme la stérilisation, le rangement ou la sortie de l'outil, changent la forme des images elliptiques d'une chirurgie à l'autre. Les manipulations de l'endoscope peuvent même causer des évolutions sur l'ellipse au cours d'une même chirurgie comme le montre la figure 4.7 (déplacement de son centre et changements de forme).

Comme la durée et le nombre de vidéos est assez élevé et que les séquences sont très bruitées (lumière pas réglée, déformations, outil à l'extérieur de l'œil...), nous avons décidé de sélectionner, pour chacune des 23 vidéos, 2 extraits. Leur durée est comprise entre 10 et 20 secondes et correspond à une phase de la chirurgie où l'endoscope est déjà en place dans l'œil. Leur durée permet d'éviter d'éventuels problèmes liés à l'évolution de la forme de la zone utile. Ces extraits présentent aussi la caractéristique de contenir des déplacements, soit de l'endoscope lui-même, soit d'autres outils également présents à l'image et interagissant avec la chambre postérieure, soit une combinaison des deux. Pour ces 46 extraits, le diamètre horizontal moyen de l'ellipse est de 336.5 pixels, avec un écart type de 15 pixels. Le diamètre vertical moyen de l'ellipse est de 370 pixels avec un écart type de 14.9.

Le taux de rafraîchissement des vidéos étant de 25 images par seconde, nous disposons d'environ 23 000 images de déplacements dans la chambre postérieure de l'œil. Pour cette étude, nous souhaitons réaliser des cartes pluri-images de la rétine grâce, notamment, à des réseaux de neurones à apprentissage fortement supervisé. Pour ce faire il nous faut donc des bases de données accompagnées des vérités terrain des déplacements inter-images. Le problème est que cette base de données de déplacements n'est pas annotée et que le faire manuellement serait imprécis et extrêmement chronophage. En effet, apparier 5 à 10 points d'intérêt par paire d'images n'est pas envisageable pour une telle quantité de données. De plus, obtenir des champs de déplacements vérité terrain pendant l'acquisition en modifiant le dis-

positif d'acquisition est tout aussi inenvisageable. Voilà pourquoi nous avons utilisé et/ou créé d'autres bases de données dont nous connaissions précisément les champs de déplacements.

Num vidéo	Outil d'acquisition	Durée totale	Durée exploitable	Pathologie	Cerclage	Pelage	Photo-coagulation	Vitrectomie
1	Endoscope	7min25s	6min26	Glaucome			X	
2	Endoscope	31min57s	14min38s	Hémorragie			X	X
3	Endoscope	51min21s	33min5s	Traumatisme				X
4	Endoscope	21min45s	17min46s	Traumatisme				X
5	Endoscope	26min01s	21min55s	Endophtalmie				X
6	Endoscope	35min12s	26min47s	Endophtalmie				X
7	Endoscope	7min55s	6min23s	Décollement de rétine				X
8	Endoscope	29min57s	26min33s	Décollement de rétine				X
9	Endoscope	63min52s	55min38s	Décollement de rétine		X		X
10	Endoscope	61min04s	36min14s	Décollement de rétine			X	X
11	Endoscope	13min52s	11min49s	Décollement de rétine	X	X		X
12	*Endoscope*	*49min27s*	–	*Décollement de rétine*				*X*
13	Endoscope	49min44s	43min40s	Décollement de rétine		X	X	X
14	Endoscope	47min03s	41min58s	Décollement de rétine		X		X
15	Endoscope	12min09s	8min35s	Décollement de rétine	X			X
16	Endoscope	52min07s	39min40s	Décollement de rétine		X	X	X
17	Endoscope	27min44s	16min03s	Décollement de rétine			X	X
18	Endoscope	24min56s	21min41s	Décollement de rétine	X			X
19	Endoscope	32mmin38s	23min16s	Décollement de rétine		X	X	X
20	*Endoscope*	*37min26s*	*2min20s*	*Décollement de rétine*				*X*
21	Endoscope	30min45s	27min13s	Décollement de rétine				X
22	Endoscope	50min01s	21min10s	Décollement de rétine			X	X
23	Endoscope	19min37s	15min08s	Décollement de rétine		X		X
24	Endoscope	48min40	7min35s	Décollement de rétine				X
25	Endoscope	43min22s	39min02s	Décollement de rétine		X		X
26	Lampe à fente	5min4s	3min3s	–				
27	Lampe à fente	26s	26s	–				
28	Lampe à fente	1min51s	21s	–				

TABLE 4.1: Tableau récapitulatif des bases de données de vidéos.
En italique : les deux vidéos non prises en compte pour la suite de l'étude

Pour l'ensemble des raisons précédentes et surtout parce que le but final est d'estimer des déplacements entre deux images de rétines, nous choisissons de faire la suite de nos tests sur la version du réseau FlowNet Simple pré-entraînée sur Flying

Chairs et affinée sur Sliding Retinas II.

4.3 Les autres bases de données

Comme précisé dans le chapitre précédent, la base de données joue un rôle primordial dans l'entraînement d'un réseau. En revanche, il est souvent compliqué (car trop long à mettre en place) d'obtenir une base de données de cas concrets, annotés, (pour l'apprentissage supervisé) suffisamment grande et diversifiée pour réaliser un apprentissage optimal. Cette réalité est d'autant plus vraie dans le domaine des bases de données médicales. En effet, il n'est pas toujours aisé de collecter des données médicales d'une part et, d'autre part, il est encore plus difficile de pouvoir proposer des annotations précises et fiables de ces données.

Voilà pourquoi nous avons décidé d'entraîner et tester nos méthodes sur des bases de données de cas simulés dont nous connaissons parfaitement les annotations et qui sont suffisamment grandes et diverses pour entraîner nos CNN. Ces bases ne sont pas nécessairement liées au domaine médical. Le détail de chacune est réalisé dans la section suivante. La première partie porte sur la base Flying Chairs [1]. La seconde porte sur les bases Sliding Retinas que nous avons développée. Enfin, la partie trois traite de la base KITTI de [115].

4.3.1 Flying Chairs

Comme évoqué précédemment, la base Flying Chairs a été créée à partir d'images de Flickr, servant de fonds, et de chaises modélisées en 3D par Aubry et al. [116], mises au premier plan. La raison de la création de Flying Chairs est que les bases de données de paires d'images avec vérités terrain des déplacements déjà existantes étaient trop petites pour l'entraînement d'un CNN à apprentissage fortement supervisé. En effet, la base de référence dans le domaine était, MPI Sintel [64] qui contient environ 1 000 paires d'images ainsi que la vérité terrain des déplacements. Flying Chairs en compte plus de 20 000 (22 872).

Les images d'arrière-plan originales sont au format 1024×768 pixels. Au premier plan de ces images est ajouté un nombre de chaises fixé aléatoirement selon une loi uniforme entre $[16\,;\,24]$. Les types, angles de vue et emplacements initiaux des chaises sont également fixés de manière aléatoire et uniforme. Les tailles des chaises (en pixels) sont définies à partir d'une gaussienne de moyenne 200 et d'écart type 200, puis bornée entre 50 et 640. Ces nouvelles images sont appelées les grandes images initiales (I_{G1}).

Pour générer la deuxième image de la paire, I_{G2}, ainsi que la vérité terrain du déplacement, Dosovitskiy et al. appliquent des déplacements aléatoires aux chaises et à l'arrière-plan. Ces déplacements correspondent à des mouvements paramétriques incluant changements d'échelle, rotations et de translations. Ces trois paramètres sont fixés aléatoirement de manière à suivre la distribution des déplacements de la base MPI Sintel (illustrée à la Figure 4.8). Le tableau 4.2 montre les bornes de chacun des paramètres.

Chaque image 1024×768 (I_{G1} et I_{G2}) est ensuite découpée en 4, donnant 4 images de taille 512×384 pixels (4 I_1 et 4 I_2) . Les informations sur les déplacements

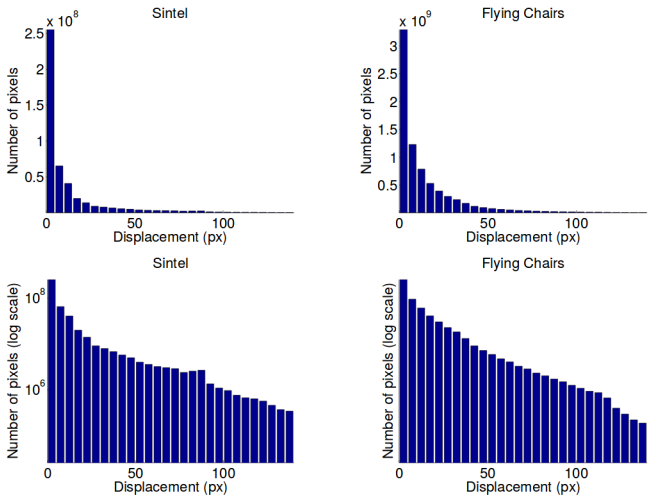


FIGURE 4.8: Compraison de la répartition des déplacements entre les bases Sintel et Flying Chairs.
Version modifiée de [1]

Déplacement	Min	Max
Translation fond (px)	-40	+40
Rotation fond (°)	-10	+10
Zoom fond	+0.93	+1.07
Translation chaises (px)	-120	+120
Rotation chaises (°)	-30	+30
Zoom chaises	+0.8	+1.2

TABLE 4.2: Amplitudes des différents déplacements des Flying Chairs [1]

entre chaque paire sont stockées sous forme de cartes denses de flux optique. Elles sont aussi de taille 512 × 384 pixels (Figure 4.9).

4.3.2 Sliding Retinas I et II

Ces deux bases de données ont été créées dans le cadre de cette étude pour palier à l'absence de grandes bases de données annotées de ce type dans le domaine médical et plus précisément ophtalmologique. Elles sont inspirées de la base Flying Chairs. En effet, il s'agit encore de déplacements générés artificiellement entre paires d'images. Les images sont également de taille 512 × 384 pixels et la vérité terrain des déplacements associée à chaque paire d'images est stockée sous forme de cartes de flux optique. Enfin, chacune des bases est composée de 23 000 paires d'images et cartes de flux optique associées. Elles ont été créées pour palier au problème d'absence de vérité terrain de nos données endoscopique, nécessaire à l'entraînement d'un CNN à apprentissage fortement supervisé. Pour tester l'efficacité de tels réseaux sur la base de chirurgies endoscopiques, il est donc important que ces nouvelles bases soient aussi proches que possible en terme de contenu de la base de chirurgies

(a) (b) (c)

FIGURE 4.9: Image initiale, image déplacée et carte de flux optique correspondant au déplacement. Base Flying Chairs. a. I_{ini} - b. I_{dep} - c. Carte de flux optique

endoscopiques.

Comme leur nom l'indique, les images qui constituent Sliding Retinas I et II représentent des rétines et plus précisément des parties de rétines, proches de celles que l'on peut trouver dans la base issue des chirurgies décrites dans la section précédente. Afin de les constituer, nous avons sélectionné aléatoirement 23 000 images de fonds d'œil dans la base Kaggle's Diabetic Retinopathy Detection (*https ://www.kaggle.com/c/diabetic-retinopathy-detection*) (Figure 4.10). Cette base a vu le jour dans le cadre d'un concours pour détecter les différents stades de rétinopathie diabétique. Elle contient plus de 35 000 images de fond d'œil de tailles variables.

FIGURE 4.10: Image Kaggle.
Image extraite du site : https ://www.kaggle.com

Pour chacun des 23 000 fonds (I_{fond}), une première imagette de taille 512 × 384 pixels est sélectionnée. Elle est désignée comme l'image initiale (I_1) de la future paire d'images. Afin de nous assurer que les deux images de la paire soient entièrement incluses dans I_{fond} et contiennent de l'information, le centre de I_1 est sélectionné

aléatoirement de manière à être éloigné au maximum de 20 pourcents du centre de I_{fond} (Comme l'illustre le schéma figure 4.11).

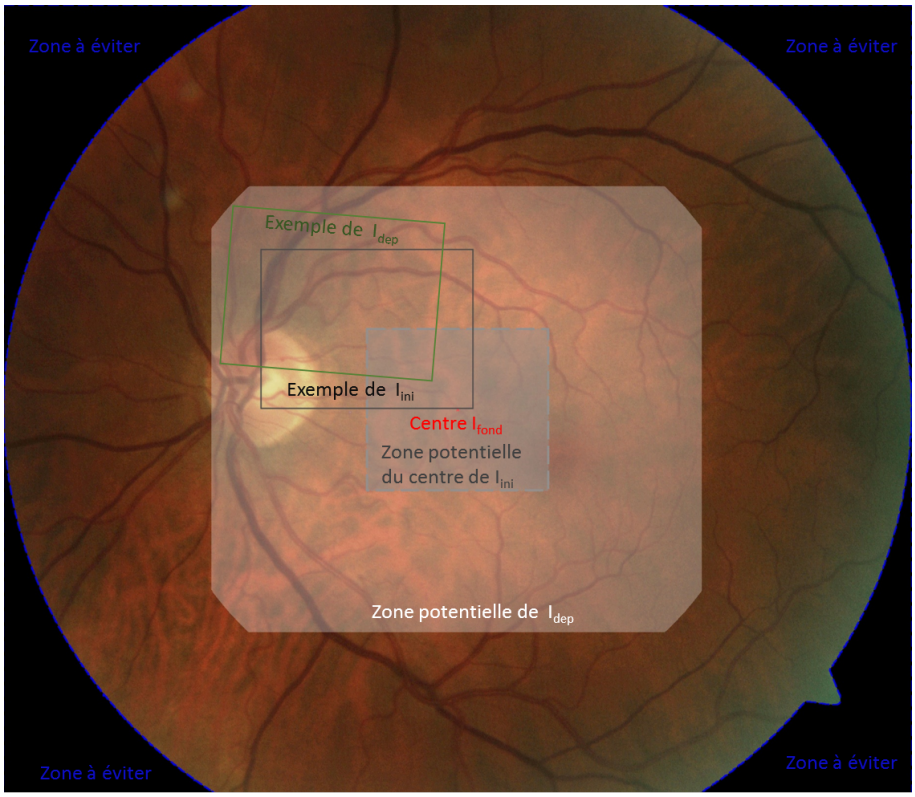
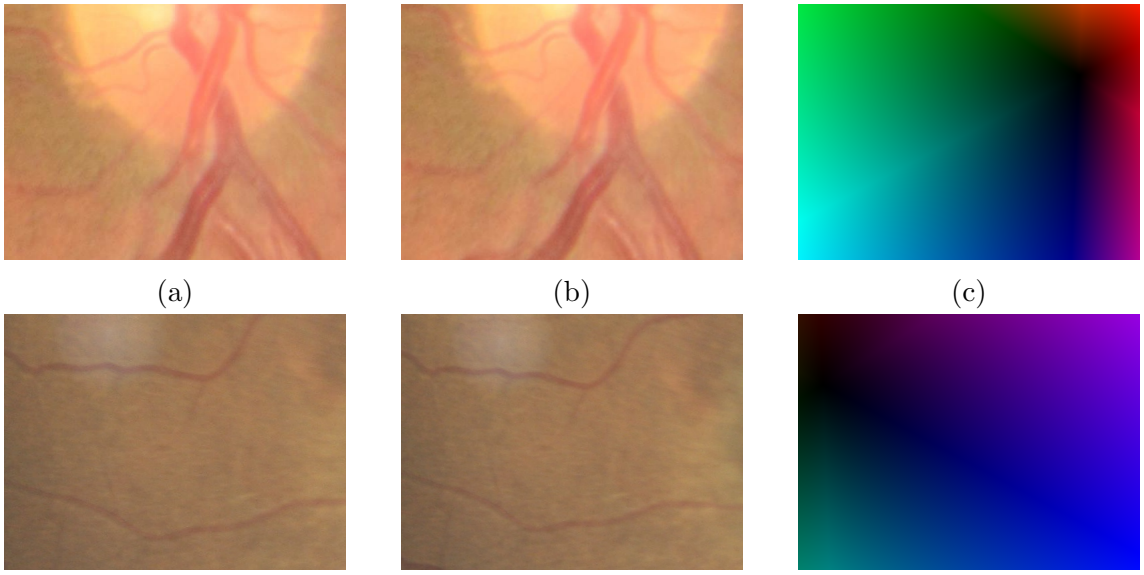


FIGURE 4.11: Schéma de création de la base Sliding Retinas.

Pour obtenir la seconde imagette I_2 nous commençons par créer le déplacement. Pour la base Sliding Retinas I, celui-ci est composé d'une translation (T) en x et en y, et d'une rotation (R). Pour la base Sliding Retinas II, il est en plus composé d'un changement d'échelle (Z), afin d'exploiter une base de données plus réaliste vis-à-vis des déplacements perçus au sein des séquences d'endoscopie. La contribution de chaque paramètre est ensuite ajoutée pour former un déplacement global comme le montre l'équation (4.1). Celui-ci est appliqué à I_{fond}. Ce nouvel I_{fond} est ensuite modifié de manière à redevenir une image parfaitement rectangulaire aux mêmes dimensions que le I_{fond} original. En se plaçant aux mêmes coordonnées que celles de l'imagette initiale, nous obtenons donc I_2 (exemple d'images et carte de flux en Figure 4.12).

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = T + R.Z \quad \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \tag{4.1}$$

Le tableau 4.3 regroupe les amplitudes de chaque paramètre constituant les déplacements de Sliding Retinas I et II. En gardant en tête que nous souhaitons une méthode capable de traiter des informations en temps réels sans pour autant bénéficier des machines les plus puissantes, nous avons considéré qu'une cadence de 5 images par seconde pouvait être envisageable. La fréquence d'acquisition des vidéos



(a) (b) (c)

(d) (e) (f)

FIGURE 4.12: Image initiale, image déplacée et carte de flux optique correspondant au déplacement. Base Sliding Retinas I. a. I_1 - b. I_2 - c. Carte de flux optique. Base Sliding Retinas II. d. I_1 - e. I_2 - f. Carte de flux optique

de chirurgie étant de 25 images par seconde, nous avons estimé les déplacements pour des paires espacées de 5 images. Afin d'ajuster les paramètres des transformations paramétriques, nous avons sélectionné une quinzaine de paires d'images sur l'ensemble des extraits vidéos d'endoscopie oculaire. Pour ces extraits de séquences réelles, il est ressorti que le déplacement moyen était de 24.9 pixels. Nous remarquons également que les déplacements semblent plus importants et variables sur l'axe horizontal que vertical et les rotations comme les changements d'échelles sont progressifs et assez faibles.

A partir de ces observations nous construisons les bases de déplacement simulés : Sliding Retinas I et II. Chacun des paramètres de déplacement d'une paire est sélectionné aléatoirement selon une loi uniforme dont les bornes sont celles du tableau 4.3. Elles correspondent approximativement aux déplacements maximaux observés sur les extraits de chirurgie étudiés. Le déplacement moyen obtenu est de 24.6 pixels pour Sliding Retinas I et 25.5 pixels pour Sliding Retinas II.

	Sliding Retinas I		Sliding Retinas II	
Déplacement	Min	Max	Min	Max
Translation en x (px)	-35	+35	-35	+35
Translation en y (px)	-26	+26	-26	+26
Rotation fond ($^\circ$)	-5	+5	-5	+5
Zoom fond	-	-	+0.9	+1.1

TABLE 4.3: Amplitudes des différents déplacements des Sliding Retinas I et II.

4.3.3 KITTI

KITTI vision Benchmark Suite [115] est un projet qui vise à constituer un ensemble de bases de données avec vérités terrains de scènes extérieures du monde réel. Pour ce faire, les chercheurs ont équipé une voiture de deux caméras couleur Flea2 14S3C de la marque Point Grey et deux caméras en niveaux de gris Flea2 14S3M également de la marque Point Grey. Ce dispositif leur a permis de constituer des bases de données pour l'étude du flux optique 2D et 3D, l'estimation de la profondeur de scènes, le suivi d'objets 2D et 3D et enfin le suivi odométrique. Pour l'ensemble de ces bases, les vérités terrain sont fournies par un scanner laser de marque Velodyne modèle HDL-64E (ou LiDAR pour Light Detections And Ranging) et un système GPS de marque OXTS modèle RT 3003 tous deux fixés et synchronisés avec le véhicule (Figure 4.13). La fréquence d'acquisition du LiDAR étant de 10 images par secondes, les caméras sont elles aussi fixées à un déclenchement de 10 images par seconde. La taille des images (1392 × 512 pixels) est également définie de manière à être compatible avec les sorties du LiDAR.

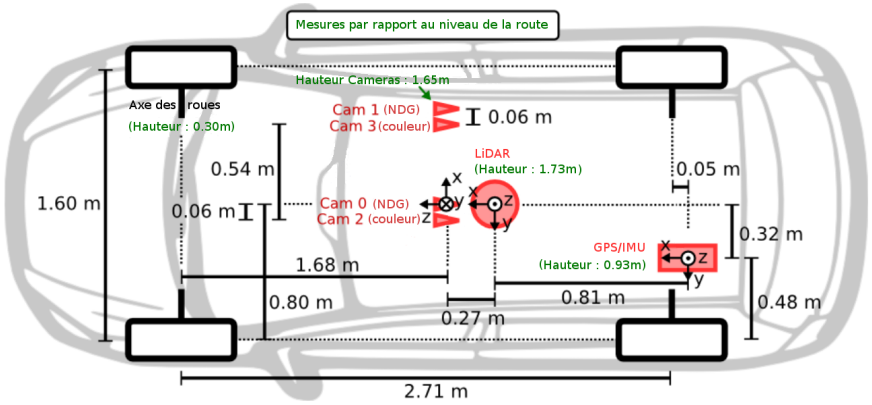


FIGURE 4.13: Schéma du système d'acquisition pour la base KITTI.
Image extraite du site : http ://www.cvlibs.net et modifiée

Les données sont acquises sur plusieurs itinéraires dans la ville de Karlsrhue (en Allemagne). Il en résulte 155 vidéos, soit plus de 48 000 images. Ces vidéos sont réparties selon 6 catégories. Dans les 4 premières catégories, le véhicule est en mouvement et pour les deux dernières, il est statique. Elles sont regroupées dans le tableau 4.4. Un exemple d'image de chaque catégorie est proposé en Figure 4.14.

Catégories	Ville	Zone Résidentielle	Route	Campus	Personne	Calibrage
Nombre de vidéos	28	21	12	10	79	5
Nombre d'images	8477	28404	5865	1308	4651	102

TABLE 4.4: Répartition des données dans la base KITTI

En plus des signaux des quatre caméras, de celui du LiDAR et des coordonnées GPS, cette base rend aussi disponible la date et l'heure de chaque prise de vue, la vitesse du véhicule ainsi que son accélération, l'altitude du matériel et son orientation

(a) (b) (c) (d) (e) (f)

FIGURE 4.14: Six images de la base KITTI. a. Image de la catégorie "City" - b. Image de la catégorie "Road" - c. Image de la catégorie "Residential" - d. Image de la catégorie "Campus" - e. Image de la catégorie "Person" - f. Image de la catégorie "Calibration"

et enfin les paramètres de calibrage de chaque élément du système. De plus, il est possible d'obtenir pour chaque image, le nombre de voitures, camions, tramways, piétons et cyclistes présents. L'acquisition et l'annotation de telles séquences donnent à cette base une complexité permettant de la comparer à des cas concrets de la vie réelle. En effet, la plupart des bases de données de référence dans le domaine sont des bases de données où les déplacements et/ou les scènes sont simulés.

4.4 Bilan

Dans ce chapitre nous avons vu un historique sur les deux systèmes d'acquisitions d'images étudiées dans cette thèse, à savoir la lampe à fente et l'endoscope oculaire. De plus, nous avons constaté que pour effectuer des expériences sur l'estimation des déplacements dans ces vidéos, il fallait préalablement faire un tri. En effet certains extraits ne sont pas exploitables, car ils ne présentent pas des images de rétines ou tout simplement parce qu'ils sont de trop mauvaises qualités. Notre objectif principal étant d'estimer des déplacements entre deux images de rétines, nous avons décidé de faire ce tri manuellement et de garder l'idée de le faire automatiquement comme piste pour de futures recherches. Enfin, ce tri pourra se révéler utile pour jouer le rôle de vérité terrain et comparer les performances des méthodes qui seront développées.

A partir de ces deux types d'acquisition vidéos nous construisons deux bases de données. Les 14 000 et 23 000 images qui composent respectivement la base lampe à fente et la base endoscope oculaire sont des données brutes et ne sont pas annotées. Nous n'avons donc aucune information quant à la vérité terrain des déplacements

qui se produisent dans ces séquences d'images. Or, pour certaines méthodes (CNN à apprentissage supervisé) d'estimations de déplacements, il nous faut cette information. Dans le domaine ophtalmologique et plus généralement dans le domaine médical, il est difficile de construire et de trouver de grandes bases de données annotées présentant des déplacements. Voilà pourquoi nous avons décidé de travailler avec des bases plus généralistes (comme Flying Chairs et KITTI) et également de construire nos propres bases de données de déplacement d'images rétiniennes (Sliding Retinas I et II). Les expériences (et leurs résultats) réalisées à partir des bases de données abordées dans ce chapitre font l'objet des deux chapitres suivants. Le chapitre 5 porte sur les méthodes dites classiques et les méthodes basées CNN sont proposées dans le chapitre 6.

5
Méthodes classiques

Ce chapitre se consacre à l'utilisation de méthodes dites classiques pour estimer les déplacements dans des vidéos acquises à la lampe à fente ou à l'endoscope oculaire. Pour certaines méthodes, la transformation estimée se traduit par une équation mathématique. En prenant comme postulat que, de proche en proche, nous pouvons considérer les images comme représentant des scènes en 2 dimensions, nous avons choisi d'estimer des transformations de type homographique. Ici, une homographie se traduit par le déplacement d'un plan dans l'espace, incluant des translations et des rotations potentielles dans tous les axes et donc de gérer les déformations d'images dues à la perspective.

Ces méthodes sont qualifiées de classiques en opposition aux méthodes basées CNN du chapitre 6. On trouve, dans un premier temps, une méthode utilisant le flux optique. La seconde partie porte sur une méthode de block matching. Enfin, la troisième partie traite des méthodes utilisant la détection automatique de points d'intérêt. L'ensemble des méthodes évoquées précédemment a été implémenté et testé pour l'estimation de déplacements entre deux images d'une séquence vidéo. Elles ont toutes été testées sur les bases de données de vidéos de lampe à fente et d'endoscopie oculaire.

5.1 La méthode utilisant le flux optique

Au cours de cette thèse, une des méthodes les plus utilisées dans la littérature des dix dernières années a été utilisée. Il s'agit de la méthode d'estimation du flux optique par l'algorithme de Gunnar Farnebäck [22]. Une première partie développe brièvement les principes de la méthode et une seconde présente les résultats sur les deux modalités d'acquisition de vidéos.

5.1.1 Méthode

Il est donc très commun de trouver cette méthode dans l'état de l'art pour estimer les déplacements entre deux images. Nous avons utilisé une version de cet algorithme capable d'estimer le flux optique entre deux images en niveaux de gris. Cet algorithme commence par faire une estimation polynomiale quadratique des voisinages de chaque pixel des paires d'images. Les équations sont donc de la forme (6.1)

$$f(x) \sim x^T Ax + b^T x + c \tag{5.1}$$

où A est une matrice symétrique, b un vecteur et c un scalaire. Ces coefficients sont ajustés à partir de la méthode des moindres carrés pondérés aux valeurs du signal du voisinage.

La pondération possède deux composantes appelées certitude et applicabilité. On retrouve ces termes dans les méthodes de [117],[118] et [119] qui portent sur des convolutions normalisées et qui constituent la base du développement polynomial. La certitude est couplée aux valeurs des signaux du voisinage. Selon [22], on définit cette certitude comme étant égale à zéro pour le voisinage en dehors de l'image. En effet, il parait naturel de ne pas considérer des points se trouvant à l'extérieur de l'image pour l'estimation des paramètres.

L'applicabilité détermine les poids relatifs des points du voisinage en fonction de leur position. Généralement, on donnera plus de poids au point central et on fera diminuer les poids de manière radiale. La largeur de l'applicabilité détermine l'échelle des structures qui seront détectées par les coefficients d'expansion.

De plus, la méthode utilisée est une méthode itérative multi-échelles. A chaque étape, le déplacement est initialisé avec la valeur du déplacement de l'étape précédente. Traditionnellement, le déplacement est fixé à zéro lors de la première étape sauf si des connaissances sur les déplacements sont connues à priori. On commence à une grande échelle pour obtenir une estimation grossière des déplacements. On va ensuite la propager à l'échelle suivante et ainsi de suite, de manière à obtenir une estimation des déplacements de plus en plus précise.

Les déplacements estimés sont dits denses. C'est-à-dire que chaque pixel d'une image possède son propre déplacement. Celui-ci est proposé sous la forme d'une carte de flux optique. C'est une carte de même taille que l'image étudiée qui se présente sous la forme d'une image à deux canaux. Chaque pixel de la carte correspond au pixel de l'image ayant les mêmes coordonnées. Un canal est dédié au stockage de la composante horizontale du déplacement estimé et l'autre à la composante verticale. Ainsi, en combinant les pixels de la seconde image avec la carte de flux optique comme dans l'équation (6.2) , on est supposé retrouver la première image.

$$I_1[y, x] = I_2[y + \delta[y, x, 1], x + \delta[y, x, 0]] \tag{5.2}$$

5.1.2 Résultats

Cette sous-section présente les résultats de la méthode d'estimation de flux optique de Farnebäck obtenus sur les vidéos de lampe à fente et d'endoscopie oculaire.

Le paramétrage s'est fait suite à des estimations manuelles sur nos bases de données combinées aux recommandations de [22]. Nous faisons varier la largeur de l'applicabilité. En effet, une initialisation à une valeur de 3 pixels est recommandée ce qui semble adapté à la largeur des plus petits vaisseaux sanguins à détecter. En revanche, certains vaisseaux font jusqu'à une dizaine de pixels de largeur. Nous choisissons donc de faire varier ce paramètre entre 3 et 11 par pas de deux.

Concernant les tests sur la base acquise à la lampe à fente, les déplacements sont estimés sur 300 paires d'images issues des 3 vidéos dont nous disposons. En ce qui concerne les tests sur la base d'endoscopies oculaires, 300 autres paires d'images, issues des 23 chirurgies, ont été sélectionnées. L'efficacité de l'estimation des déplacements est mesurée de manière quantitative par la différence absolue moyenne(en niveau de gris). Celle-ci est calculée entre l'image initiale et l'image finale recalée sur l'image initiale par le déplacement estimé. Cette différence est calculée uniquement pour des zones communes aux deux images.

De plus, l'efficacité de l'estimation des déplacements est estimée visuellement à travers plusieurs moyens. Nous faisons dans un premier temps subir à l'image finale le déplacement estimé de manière à retrouver théoriquement l'image initiale. Afin de mettre en évidence les similitudes et les différences, nous calculons une image appelée damier. Celle-ci est composée alternativement d'images initiales et d'images finales recalées sur le modèle de l'image initiale. Un exemple de damier est proposé en figure 5.1. Dans un cas idéal, nous ne devrions pas être en mesure de distinguer de discontinuité entre chaque case du damier.

Le tableau 5.1 présente les valeurs moyennes des différences absolues des images finales recalées et initiales pour les différentes valeurs prises par la largeur de l'applicabilité, pour les deux bases de données vidéos. Chacun des chiffres pris séparément n'est pas significatif, mais leur comparaison permet de mettre en évidence quel paramétrage donne les estimations les plus précises.

Base \ Applicabilité	**3**	5	7	9	11
Lampe à fente	**22.10**	22.13	22.17	22.18	22.21
Endoscope	**45.28**	45.43	45.44	45.44	45.43

TABLE 5.1: Tableau récapitulatif des différences absolues moyennes pour la méthode de Farnebäck pour différentes valeurs de largeur d'applicabilité.
Les résultats sont obtenus sur 300 paires d'images. En gras : le paramétrage conservé pour la suite de la partie résultats

Le tableau 5.1 nous montre que pour les deux bases de données, c'est la largeur égale à 3 (paramétrage recommandé) qui minimise la différence absolue moyenne. Pour une même base, les valeurs sont proches ce qui laisse penser que les différences d'estimations de flux optiques sont assez minimes. De plus, on remarque que cette différence reste environ deux fois plus élevée pour la base acquise à l'endoscope que pour la base acquise à la lampe à fente. L'étude visuelle des damiers figure 5.2 vient compléter notre analyse. Cette figure propose de comparer les damiers obtenus pour chacune des valeur testée de la largeur et pour les deux bases de données.

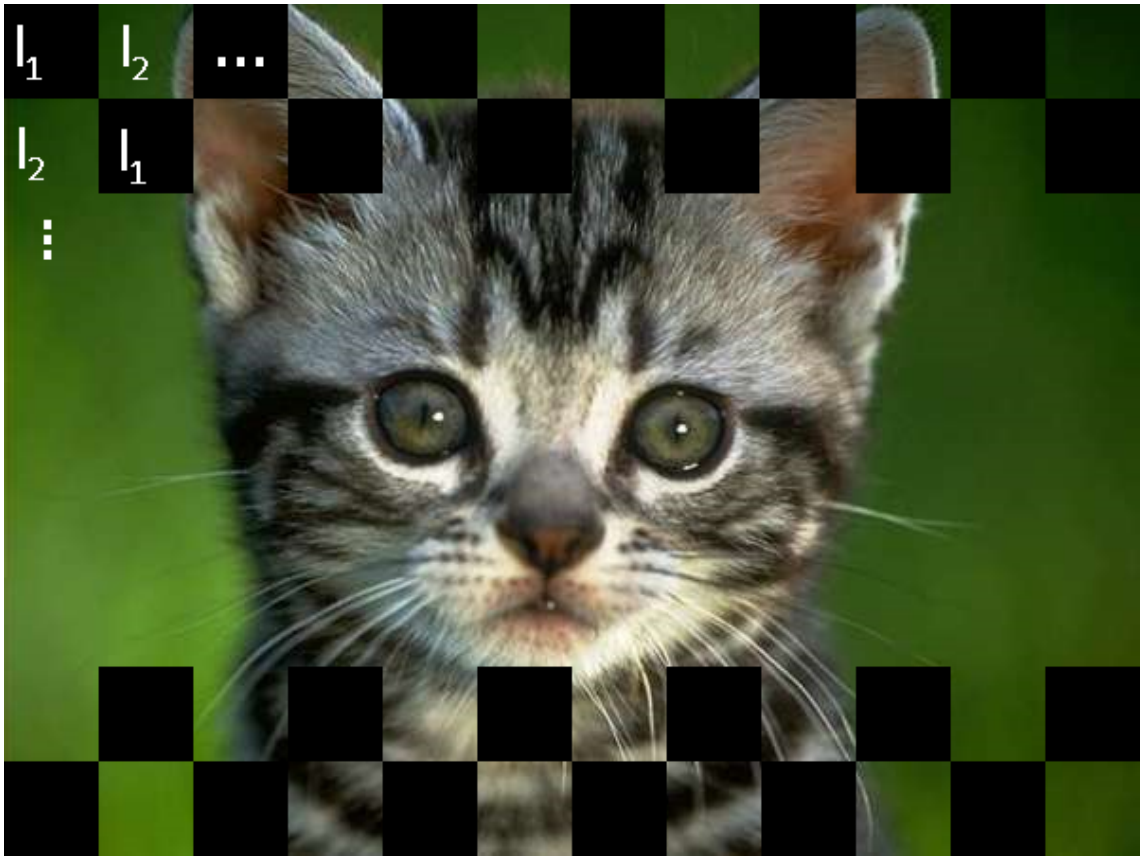
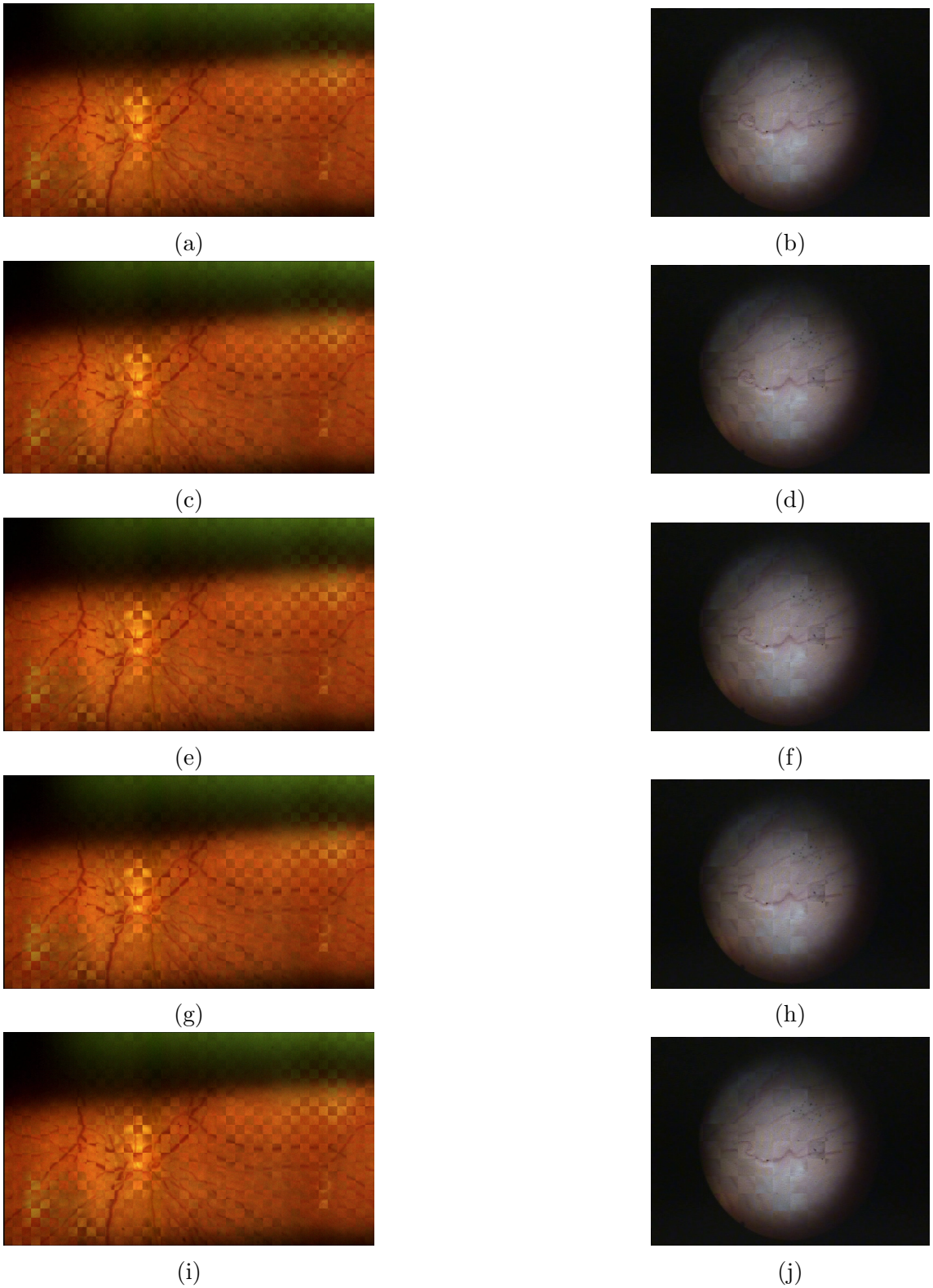
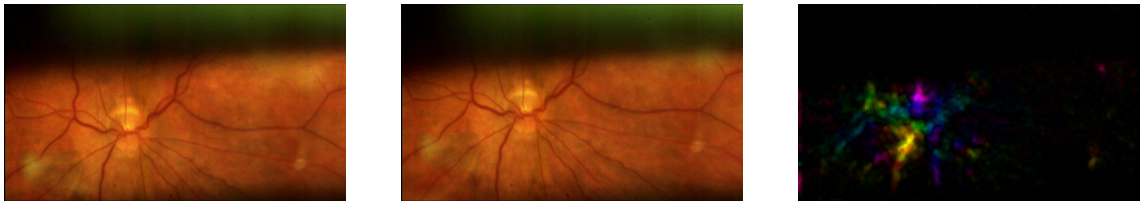


FIGURE 5.1: Exemple illustrant la composition d'un damier.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

FIGURE 5.2: Damiers obtenus pour les différentes valeurs de largeur d'applicabilité testées. a. largeur=3 base lampe à fente - b. largeur=3 base endoscope - c. largeur=5 base lampe à fente - d. largeur=5 base endoscope - e. largeur=7 base : lampe à fente - f. largeur=7 base endoscope - g. largeur=9 base lampe à fente - h. largeur=9 base endoscope - i. largeur=11 base lampe à fente - j. largeur=11 base endoscope



(a) (b) (c)

FIGURE 5.3: Exemple d'une paire d'image et de l'estimation du flux optique entre la première et la seconde image (lampe à fente). a. Première image - b. Seconde image - c. Flux optique estimé

D'après la figure 5.2, on constate que les différences des recalages en fonction de la largeur de l'applicabilité sont faibles. Ce qui confirme les résultats obtenus dans le tableau 5.1. A travers cet exemple, il semble difficile de conclure sur quelle largeur propose le recalage le plus fidèle pour la base acquise à la lampe à fente. En revanche pour l'exemple de la base d'endoscopies, il semblerait bel et bien que la largeur égale à 3 propose une estimation légèrement meilleure. Nous pouvons l'observer à travers le vaisseau sanguin central de l'image qui semble plus continu sur le damier figure 5.2b.

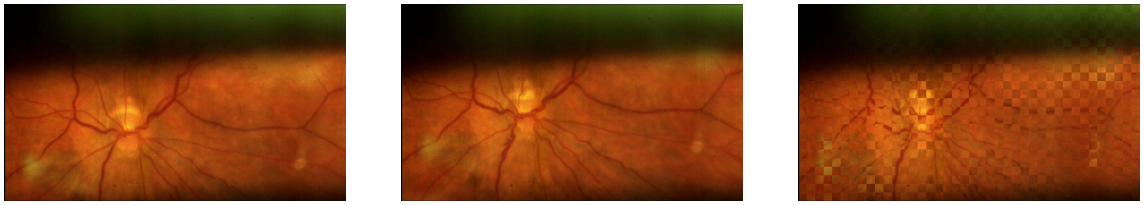
Une analyse plus détaillée des exemples d'estimation de recalage est proposée dans la suite de la section ; seuls les résultats pour une largeur de 3 sont présentés. En effet, il semblerait que ce paramétrage soit le meilleur ou au moins équivalent aux autres.

Lampe à fente

En analysant l'exemple proposé en figure 5.3 on constate qu'il y a un déplacement global entre les deux images alors que la carte de flux optique semble montrer un déplacement plus élevé dans les zones autour des gros vaisseaux sanguins et peu voire pas de déplacement dans les zones plus faiblement vascularisées.

Le damier présenté en figure 5.4 nous confirme que l'estimation du déplacement entre les deux images semble correcte par endroits, mais que, dans la majorité des cas, celle-ci n'est pas fiable. En effet, on peut observer une certaine continuité de certains vaisseaux dans le bas de l'image, mais pour beaucoup, ceux-ci sont dédoublés (comme autour de la papille ou dans la droite de l'image). Ce phénomène traduit une mauvaise estimation du flux optique.

Enfin, l'observation de l'exemple figure 5.5 qui montre la valeur absolue de la différence entre l'image finale recalée et l'image initiale confirme que l'estimation des déplacements via la méthode de Farnebäck d'estimation du flux optique n'est pas efficace dans cette modalité. Pour cet exemple, la valeur de la différence absolue moyenne est de 13.10. Cette valeur est plus faible que la valeur moyenne sur les 300 images de tests, car même sans recalage, la différence moyenne entre les deux images est faible. En effet, les textures restent globalement les mêmes et l'éclairage varie peu. La première colonne du tableau récapitulatif 5.3 reprend les valeurs de différences absolues moyennes pour la valeur de largeur égale à trois puisqu'il s'agit de la valeur donnant les meilleurs résultats pour cette méthode.



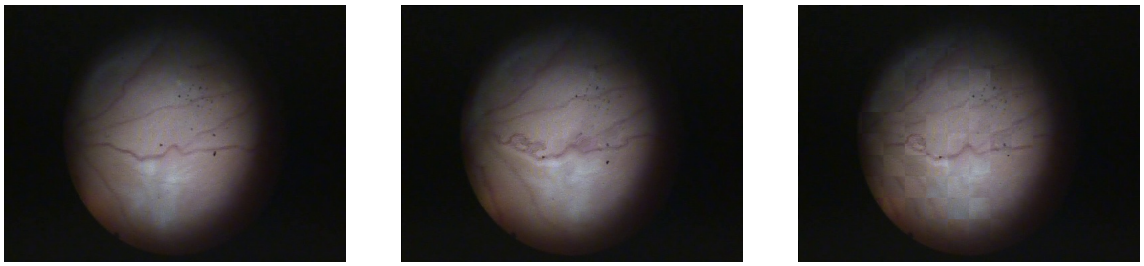
(a) (b) (c)

FIGURE 5.4: Exemple d'une paire d'image et de l'estimation du damier composé de la première et la seconde image (lampe à fente). a. Première image -b. Seconde image recalée sur la première - c. Damier

FIGURE 5.5: Exemple de différence absolue entre la seconde image recalée et la première image (lampe à fente).

(a) (b) (c)

FIGURE 5.6: Exemple d'une paire d'image et de l'estimation du flux optique entre la première et la seconde image (endoscopie). a. Première image - b. Seconde image - c. Flux optique estimé



(a) (b) (c)

FIGURE 5.7: Exemple d'une paire d'image et de l'estimation du damier composé de la première et la seconde image (endoscopie). a. Première image - b. Seconde image recalée sur la première c. Damier

Endoscopie

Comme pour le cas des tests sur les vidéos de lampe à fente, nous détaillons plus les résultats donnés par l'exemple vu en figure 5.2. Pour cet exemple, nous étudions l'estimation du flux optique, un damier ainsi que la visualisation de la valeur absolue de la différence entre l'image finale recalée et l'image initiale.

Comme dans le cas précédent, l'observation du flux optique figure 5.6 nous montre un déplacement important détecté le long d'un vaisseau sanguin principal et des déplacements beaucoup plus faibles dans le reste de l'image (zones à faible gradient). Visuellement, nous constatons que tous les pixels de la zone utile de l'image subissent un déplacement, hors ce n'est pas ce que semble traduire cette carte de flux optique.

De manière générale, les images d'endoscopie oculaire sont moins texturées que celles de lampe à fente et on le retrouve dans les exemples proposés. De ce fait, il est plus difficile d'interpréter visuellement le damier en figure 5.7. Cependant, nous pouvons retrouver dans cet exemple les mêmes observations que pour l'exemple issu de la lampe à fente. En effet, on retrouve dans toute la partie basse de l'image le dédoublement des vaisseaux sanguins, traduction d'une mauvaise estimation des déplacements. Néanmoins, nous pouvons aussi retrouver certaines zones autour du vaisseau sanguin central de l'image, pour lesquelles le recalage semble plus efficace.

Enfin, l'observation de l'exemple figure 5.8 montre que la valeur absolue des

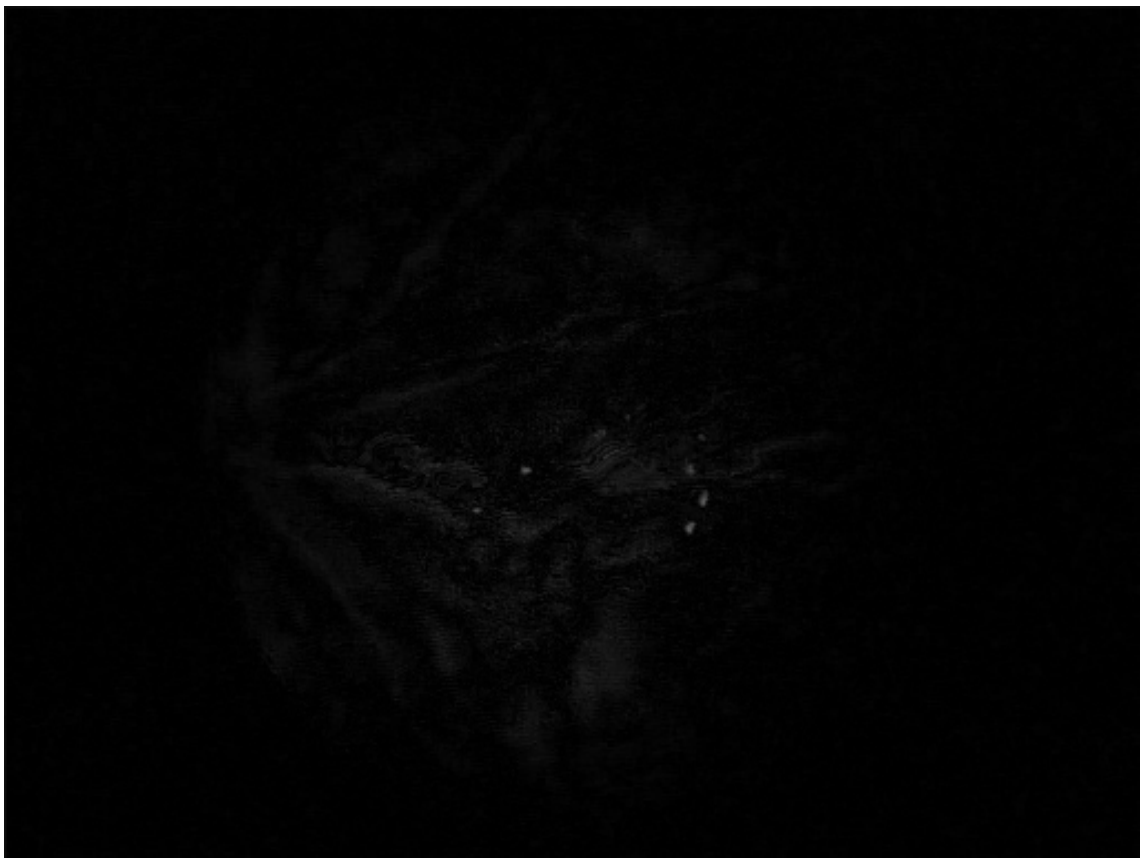


FIGURE 5.8: Exemple de différence absolue entre la seconde image recalée et la première image (endoscopie).

différences entre l'image finale recalée et l'image initiale n'est pas parfaite, mais l'intensité semble plus faible que dans le cas de l'exemple de la lampe à fente (6.87 en moyenne contre 13.10). Ce phénomène s'explique par deux raisons. La première est que pour cet exemple, le mouvement est mieux estimé que pour l'exemple de la figure 5.3. La seconde raison est que l'image est plus faiblement texturée et contrastée. Une différence entre deux zones différentes aura donc naturellement tendance à être plus faible puisqu'elles sont d'avantage homogènes.

En conclusion, nous pouvons en déduire que la méthode de Farnebäck ne semble pas adaptée à l'estimation de déplacement pour des images faiblement texturées. En effet, il est fort probable que les polynômes représentant des pixels et leur voisinage dans les zones faiblement vascularisées soient tous très proches. A l'inverse, ceux autours des zones de la papille ou des vaisseaux sanguins principaux sont plus singuliers. Cette conclusion est en partie vérifiée en observant le damier 5.7 qui montre une continuité pour certains des plus gros vaisseaux sanguins. L'observation du tableau 5.1 semble privilégier la première proposition puisque globalement, nous pouvons voir que cette différence est plus de deux fois plus grande en moyenne, pour les 300 images d'endoscopie oculaire que pour celles de lampe à fente.

Il peut être intéressant de noter que les bases ainsi que les modalités d'évaluation sont les mêmes que précédemment pour les deux autres sections développées dans la suite de ce chapitre. Ceci nous permet de comparer les méthodes entre elles dans le tableau 5.3 et d'illustrer nos résultats à travers les mêmes exemples.

5.2 La méthode utilisant le Block Matching

D'après l'état de l'art [38], les méthodes basées sur le principe de Diamond Search, évoquées dans le chapitre 2, donnent les meilleurs résultats en terme d'estimation des déplacements et de rapidité de calcul parmi les méthodes utilisant le block matching. Nous avons donc choisi de développer un algorithme basé sur la méthode Diamond Search de [39] et de la modifier pour l'adapter à notre problématique. Cette méthode se décompose en plusieurs étapes qui sont détaillées dans la partie suivante.

5.2.1 Méthode

Tout d'abord, comme dans chaque méthode de block matching, on estime les changements entre deux images. La première image est désignée comme l'image objet et la seconde est désignées comme l'image scène. Ces dénominations sont courantes dans le domaine du block matching. En revanche, par souci de lisibilité et de cohérence avec les autres sections de ce manuscrit, nous allons garder les appellations "première image ou "image initiale" pour l'image objet et "seconde image" ou "image finale" pour l'image scène.

L'image initiale est découpée en blocs de taille fixe. La plupart du temps, les blocs sont de taille 32 par 32 pixels, voilà pourquoi nous nous sommes orientés vers cette dimension. S'agissant également d'une méthode incrémentale, on commence par initialiser les vecteurs de déplacements de chaque pixel de chaque bloc à [0 ;0]. De plus, on initialise le compteur incrémental à 0. L'étape suivante consiste à comparer

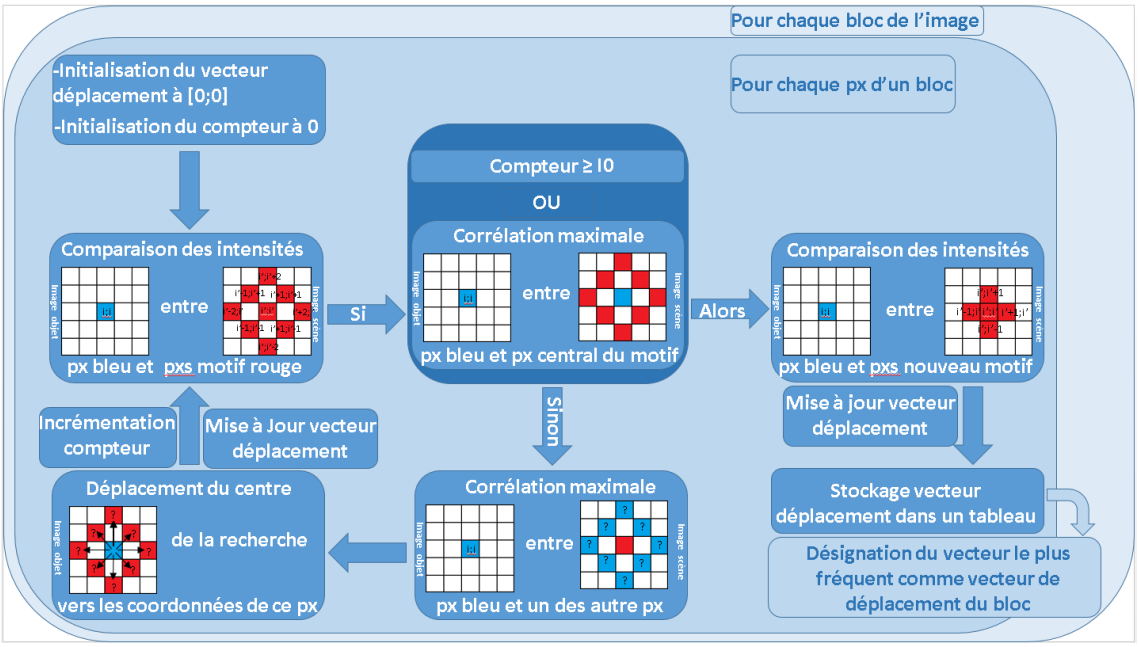


FIGURE 5.9: Schéma illustrant le fonctionnement de la méthode de block matching proposée.

les intensités des pixels des blocs de l'image initiale avec les pixels des coordonnées homologues de l'image finale. On compare aussi les intensités des pixels de l'image initiale avec les pixels adjacents de l'image finale en suivant le motif de la figure 5.9. Ce motif a la forme d'un losange, qui se dit diamond en anglais et d'où tire son nom la méthode. En fonction des résultats obtenus et de la valeur du compteur incrémental, on continue ou non la méthode. Le schéma global de la méthode est proposé en figure 5.9. Le nombre maximal d'incrément autorisé a pour but de déterminer le déplacement maximal que la méthode puisse détecter.

Historiquement, cette méthode est utilisée pour la compression de vidéos et ne cherche pas directement à estimer des déplacements. Nous l'avons détournée de son utilisation première pour obtenir une liste de points communs aux deux images. Nous prenons comme points d'intérêt les centres des 8×8 blocs centraux de l'image nous amenant ainsi à 64 points. Avec ces listes de points, nous estimons un mouvement de type homographique.

Celui-ci se présente sous la forme d'une matrice 3×3. Pour trouver la meilleure transformation, on minimise l'écart entre l'image initiale et l'image finale transformée de manière à ressembler à l'image initiale. Idéalement, on cherche les paramètres qui permettent de vérifier l'équation (6.3)

$$I_1 = H.I_2 \tag{5.3}$$

où H est la matrice de transformation, I_1 l'image initiale et I_2 l'image finale. Dans le cas où l'image obtenue n'est pas rectangulaire, celle-ci est complétée de pixels noirs de manière à avoir les mêmes dimensions que l'image scène. On peut également constater que les transformations engendrent des pixels vides dans l'image. C'est-à-dire des pixels dont la valeur n'est pas déterminée dans l'image finale une fois celle-ci

recalée. Afin de prévenir cette situation, une interpolation linéaire est appliquée.

5.2.2 Résultats

Cette sous-section présente les résultats de la méthode d'estimation des déplacements basée sur une variante de l'algorithme Diamond Search obtenus sur les vidéos de lampe à fente et d'endoscopie oculaire. Nous avons vu dans le chapitre précédent que le déplacement moyen pour la base d'endoscopie était de 24.9 pixels au vu des contraintes que nous fixions. Cette valeur est du même ordre de grandeur pour les déplacements dans la base de vidéos acquises à la lampe à fente.

Nous avons donc fait varier le nombre d'incréments entre 8 et 16, par pas de un, fixant respectivement le déplacement maximal entre 17 et 33 pixels. Les résultats de ces tests sont présentés dans le tableau 5.2 pour les deux bases de données vidéos pour 300 paires d'images.

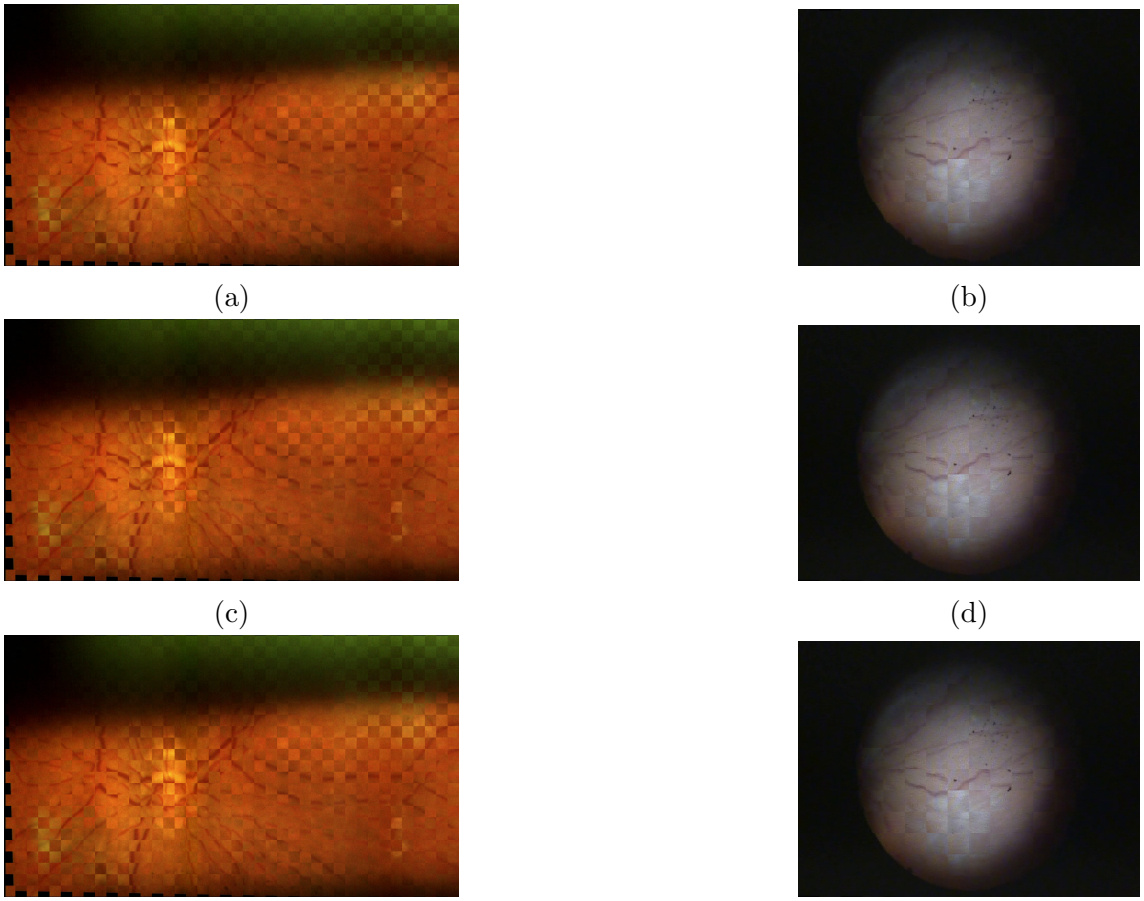
Base \ Inc Max	8	9	10	11	**12**	13	14	15	16
Lampe à fente	23.27	23.12	23.00	23.00	**23.00**	23.00	23.00	23.00	23.01
Endoscope	48.75	48.60	47.88	47.79	**47.53**	47.53	47.53	47.53	47.53

TABLE 5.2: Tableau récapitulatif des différences absolues moyennes pour la méthode Diamond Search.
Les résultats sont obtenus sur 300 paires d'images. En gras : le paramétrage conservé pour la suite de la partie résultats

Le tableau 5.2 nous montre que le nombre d'incréments minimisant la différence absolue moyenne est obtenu à partie de 10 pour la base de données acquise à la lampe à fente et 12 pour la base acquise à l'endoscope oculaire. Encore une fois, pour chaque base, l'écart semble faible pour les différentes valeurs de différence absolue moyenne. On remarque même que cette valeur stagne une fois le minimum obtenu, ce qui semble cohérent. En effet, une fois la valeur optimale d'incrément trouvée, il semble logique que même en offrant à l'algorithme la possibilité de répéter l'opération de recherche de minimum n fois supplémentaires, celui-ci va garder en mémoire l'étape qui le minimise. Globalement, les valeurs sont proches des valeurs obtenues avec le test précédent ce qui laisse penser que les estimations des déplacements ne sont pas bonnes.

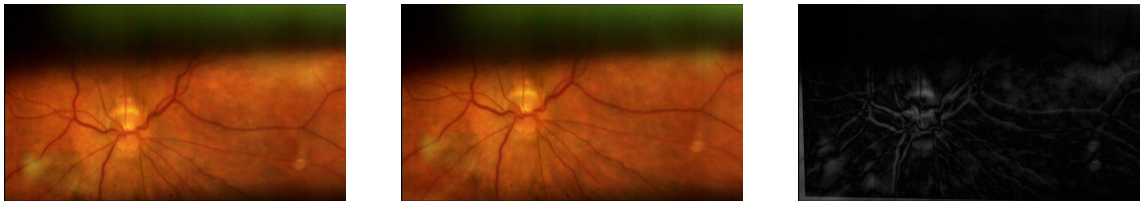
La figure 5.10 présente les damiers pour les valeurs extrêmes prises pour le nombre d'incréments à savoir 8 et 16. On peut également y trouver les damiers obtenus pour la valeur minimisant l'erreur absolue moyenne pour les deux bases à savoir 12. Pour les deux bases, nous ne constatons pas de différences sur les damiers pour les différentes valeurs présentées.

Comme dans la partie dédiée à la méthode de Farnebäck, la suite de la partie ne présentera que les résultats obtenus avec la valeur optimale pour le nombre d'incréments maximal, à savoir 12.



(a) (b) (c) (d) (e) (f)

FIGURE 5.10: Damiers obtenus pour différentes valeurs de l'incrément maximal. a. inc max=8 base lampe à fente - b. inc max=8 base endoscope - c. inc max=12 base lampe à fente - d. inc max=12 base endoscope - e. inc max=16 base : lampe à fente - f. inc max=16 base endoscope.



(a) (b) (c)

FIGURE 5.11: Exemple d'une paire d'image et de la différence absolue entre la première et la seconde image (lampe à fente). a. Première image - b. Seconde image - c. Différence absolue des deux images

(a) (b) (c)

FIGURE 5.12: Exemple d'une paire d'image et de l'estimation du damier composé de la première et la seconde image (lampe à fente). a. Première image - b. Seconde image recalée sur la première - c. Damier

Lampe à fente

L'observation de l'exemple figure 5.11, correspondant à la valeur absolue de la différence entre la seconde image recalée et la première image, semble présager que l'estimation des déplacements via la méthode de block matching n'est pas efficace pour ce cas. En effet, on retrouve assez nettement un dédoublement de la papille ainsi que des vaisseaux sanguins principaux. Cette image est sensiblement la même que celle observée dans la partie précédente. Là où, pour cet exemple, on constatait une différence moyenne de 13.10 avec la méthode de Farnebäck, on trouve ici 12.76. La différence est légèrement plus faible, mais semble trop faible pour pouvoir être observée.

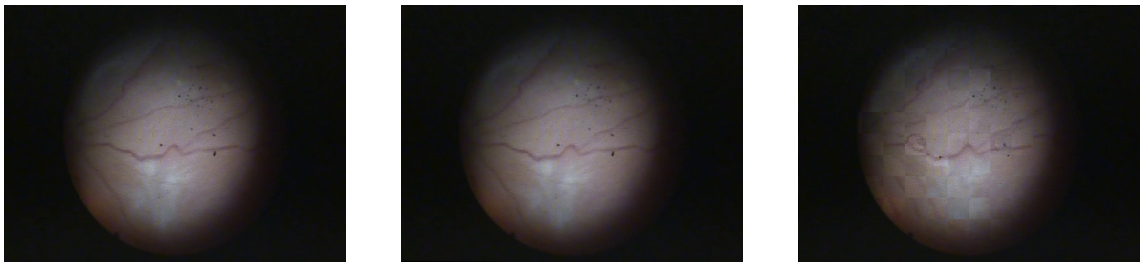
De manière plus globale, la valeur en gras dans le tableau 5.2 nous montre que nous avons une différence absolue moyenne globale sur les 300 cas de 23.0 contre 22.1 avec la méthode de Farnebäck. Des résultats assez proches qui traduisent une mauvaise estimation des déplacements.

Le damier présenté en figure 5.12 confirme cette constatation. En effet, là où nous pouvions observer par endroit une certaine continuité de quelques vaisseaux sanguins avec la méthode précédente, avec celle-ci tous les vaisseaux sont dédoublés.

Pour expliquer ce dysfonctionnement, la même hypothèse que précédemment peut être évoquée. A savoir que les images ne sont probablement pas assez texturées pour qu'un tel algorithme, pourtant reconnu et utilisé dans la littérature, puisse être exploitable dans le cadre de notre problématique.

(a) (b) (c)

FIGURE 5.13: Exemple d'une paire d'image et de la différence absolue entre la première et la seconde image (endoscopie). a. Première image - b. Seconde image - c. Différence absolue des deux images



(a) (b) (c)

FIGURE 5.14: Exemple d'une paire d'image et de l'estimation du damier composé de la première et la seconde image (endoscopie). a. Première image - b. Seconde image recalée sur la première - c. Damier

Endoscopie

Comme les exemples des figures 5.13 et 5.14 le montrent, cette méthode ne semble également pas pouvoir s'appliquer sur la base des vidéos d'endoscopie oculaire. En effet, comme ces vidéos sont encore plus faiblement texturées et contrastées que celles de la base des vidéos de lampe à fente, il parait logique d'obtenir de tels résultats.

Enfin, pour appuyer cette hypothèse nous pouvons voir que dans la globalité, là encore, l'erreur moyenne globale pour la base d'endoscopie est, comme le montre le tableau 5.2, plus de deux fois supérieure à celle obtenue pour la base vidéos acquises à la lampe à fente.

5.3 Les méthodes utilisant des points d'intérêt

Ces méthodes se rapprochent par certains aspects de la méthode précédente, car là encore, on cherche à estimer un déplacement à partir de points caractéristiques de chaque image (méthode parcimonieuse). La différence est que dans la méthode précédente, les points sélectionnés sont à des coordonnées fixes tandis que les méthodes basées points d'intérêt cherchent à détecter automatiquement des points saillants.

Comme évoqué dans le chapitre 2, ces méthodes sont décomposées en trois étapes principales à savoir la détection des points d'intérêt, l'appariement de ces points entre

les deux images et enfin l'estimation d'un déplacement à partir des correspondances de points.

SIFT

Le premier algorithme de détection de points d'intérêt de la littérature utilisé est extrait de l'algorithme SIFT [45]. Dans cette méthode, ce que nous appelons points d'intérêt se retrouve sous l'appellation de descripteurs d'images. Comme leurs noms l'indiquent, ces descripteurs correspondent à de l'information détectée localement sur une image et qui est supposée contenir certaines caractéristiques visuelles de l'image. La caractérisation est supposée être robuste aux changements d'échelle, à l'angle d'observation et à l'exposition de l'image. Ainsi, avec cette approche, deux images d'un même objet sont supposées avoir les mêmes descripteurs.

Un point d'intérêt est défini par ces coordonnées sur l'image ainsi qu'un facteur d'échelle (σ). Le travail de recherche des points va donc se faire sur différentes échelles. Une première étape consiste à convoluer l'image (I) par un filtre de Gauss (G) de largeur σ (voir équation (6.4)). Cette étape a pour effet de lisser l'image et faire disparaître les informations de plus petites tailles que le rayon σ de la gaussienne, donnant une nouvelle image (L). La différence entre l'image originale et l'image ayant subi la convolution permet de mettre évidence ces détails.

$$L(x, y, \sigma) \; = \; G(x, y, \sigma) \; * \; I(x, y) \tag{5.4}$$

Afin de rendre ce procédé plus efficace, la convolution peut être effectuée à différentes échelles, mettant ainsi en évidence des détails plus ou moins grands. L'équation (5.5) est une généralisation pour différentes échelles de l'opération permettant de mettre en évidence les détails où k est un facteur d'échelle.

$$D(x, y, \sigma) \; = \; L \; (x, y, k\sigma) - L(x, y, \sigma) \tag{5.5}$$

SURF

La seconde méthode est plus récente et est supposée être une version plus robuste et plus rapide de la méthode SIFT. Il s'agit de la méthode SURF [46]. Là encore, il s'agit d'un algorithme de détection de points d'intérêt. Elle est basée sur l'analyse des réponses de l'image à différentes ondelettes de Haar.

Là où SIFT convolue l'image par un filtre de Gauss, SURF va le faire avec un filtre de forme rectangulaire et l'intégrale de l'image rendant l'opération plus rapide. En revanche, on retrouve ce même travail de comparaison d'image à différentes échelles dans les deux méthodes.

Une fois un point saillant détecté, une caractérisation de celui-ci va se faire en analysant son voisinage et permettant ainsi une certaine robustesse aux changements d'échelle et surtout aux changements d'orientation. Le voisinage est de forme carrée et de taille variable. Il est centré sur le point saillant. Le voisinage, aussi appelé région d'intérêt est divisé en sous-régions. Pour chacune d'elles, on extrait les réponses aux ondelettes de Haar pour certains points régulièrement espacés de la sous-région. Pour permettre une plus grande robustesse aux bruits et aux déformations, ces réponses sont pondérées par une gaussienne.

Appariement
La méthode qui a été utilisée a pour but de trouver les meilleures correspondances parmi deux ensembles de données et est basée sur la méthode RANSAC (pour RANdom SAmple Consensus) [48]. Comme son nom l'indique elle réalise cette opération en sélectionnant itérativement et aléatoirement un sous-ensemble de points dans chacun des deux ensembles. Elle considère arbitrairement que les correspondances entre ces points sont les bonnes puis teste cette hypothèse à travers un algorithme décomposable en cinq étapes principales.

Dans la première étape, les paramètres du modèle s'ajustent pour faire en sorte que celui-ci vérifie bel et bien l'hypothèse de départ. Dans un second temps, le reste des données est testé sur le modèle établi en première étape. Si une nouvelle paire de points correspond au modèle, alors on incrémente un compteur et les données sont considérées comme pertinentes. Si ce compteur passe un certain seuil alors le modèle est considéré comme correct (c'est la troisième étape). Dans la quatrième étape, le modèle est mis à jour en lui ajoutant les nouvelles données pertinentes. Dans la dernière phase, on ré-évalue le modèle en estimant les erreurs entre celui-ci et les données pertinentes.

Ces cinq étapes sont répétées un certain nombre de fois. A l'issue de chaque itération, trois possibilités sont envisageables. Si trop peu de points correspondent au modèle, l'algorithme s'arrête en étape 2. Si suffisamment de points sont en adéquation avec le modèle, il y a un calcul d'erreur. Si l'erreur est plus faible que celle du meilleur modèle précédent alors ce nouveau modèle devient la référence et sinon le nouveau modèle est rejeté.

Estimation du déplacement
Comme précisé en introduction du chapitre, dans notre cas, le modèle recherché par la méthode est une homographie. Mathématiquement, elle se définit comme étant une transformation géométrique linéaire entre deux plans projectifs. Pour résoudre une équation homographique, il faut au minimum quatre paires de points. Il s'agit de résoudre un système d'équations à 8 inconnues. La transformation entre deux images par homographie se traduit comme dans l'équation (5.6).

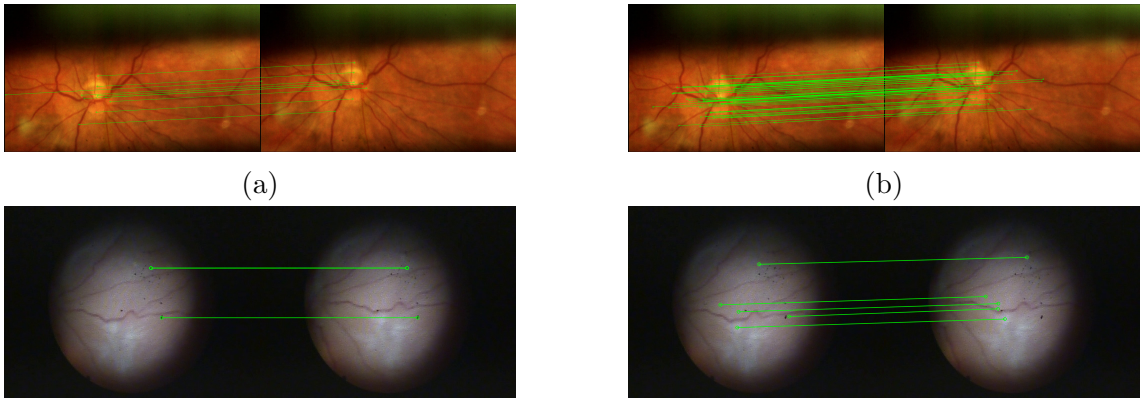
$$I_1[x,y] = I_2\left[\frac{h_{11}x + h_{21}y + h_{13}}{h_{31}x + h_{32}y + h_{33}}, \frac{h_{21}x + h_{22}y + h_{23}}{h_{31}x + h_{32}y + h_{33}}\right] \tag{5.6}$$

$$\Sigma_i \left(x'_i - \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}yi + h_{33}}\right)^2 + \left(y'_i - \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}xi + h_{32}y_i + h_{33}}\right)^2 \tag{5.7}$$

Le calcul de la meilleure homographie se fait avec la méthode de Levenberg-Marquardt [120]. Il s'agit d'une méthode de régression au sens des moindres carrés où le résultat de l'équation (5.7) doit être minimisé.

5.3.1 Résultats

Pour estimer une homographie, il faut à la fonction 4 paires de points au minimum. Nous avons donc testé les méthodes suivantes de manière à ce qu'elles



(a) (b)

(c) (d)

FIGURE 5.15: Exemple des points d'intérêt détecté automatique par méthode SIFT et SURF pour des modalités d'acquisitions différentes. a. Exemple de points d'intérêt obtenus par méthode SIFT pour une acquisition à la lampe à fente - b. Exemple de points d'intérêt obtenus par méthode SURF pour une acquisition à la lampe à fente - c. Exemple de points d'intérêt obtenus par méthode SIFT pour une acquisition à l'endoscope - d. Exemple de points d'intérêt obtenus par méthode SURF pour une acquisition à l'endoscope

détectent puis conservent les 4 paires de points les plus probables en terme d'appariement. De plus, cette méthode recommande, pour un fonctionnement optimal, un nombre de 10 paires de points. Nous avons donc également testé les méthodes pour qu'elles détectent et conservent 10 paires de points. Dans un premier temps, les résultats de la détection de points basée sur la méthode SIFT sont présentés pour (4 et 10 points d'intérêt), puis dans un second temps, les résultats obtenus avec la méthode SURF.

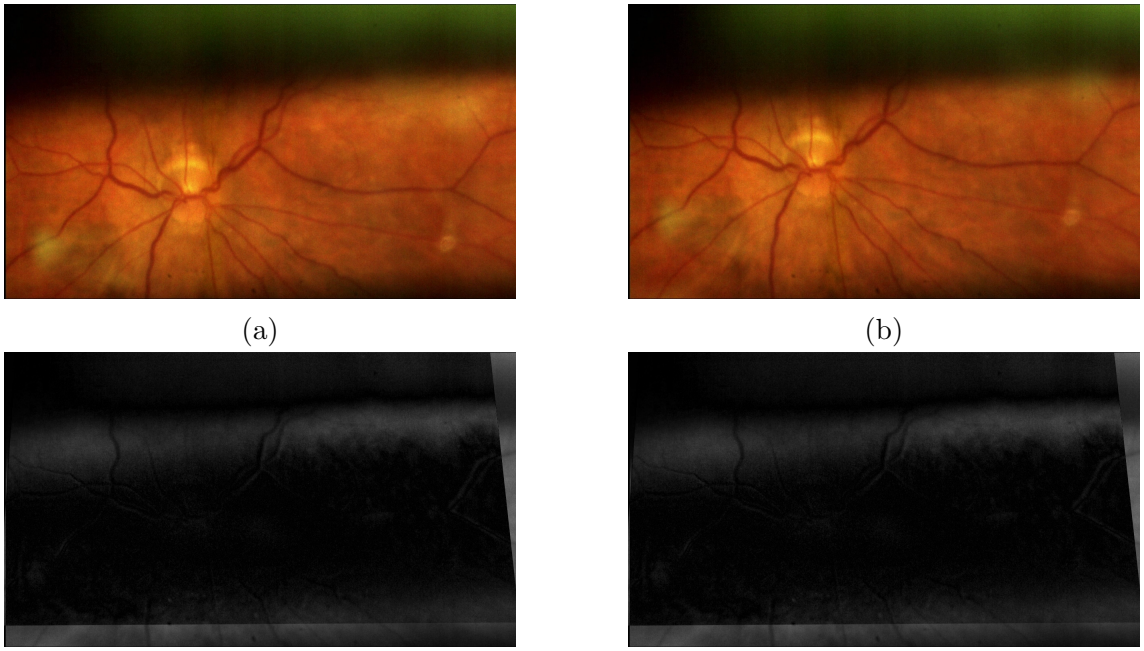
Lampe à fente

SIFT

Dans le cadre des vidéos acquises par lampe à fente, la méthode SIFT a systématiquement réussi à trouver des points d'intérêt sur les paires d'images, qui ont pu être mis en corrélation par la suite. Les points d'intérêt détectés, ainsi que leur mise en correspondance est d'ailleurs proposée en figure 5.15 pour les paires d'images prises en exemple depuis le début de ce chapitre.

Dans l'exemple de la figure 5.16 on voit que la différence entre les deux images semble plus faible qu'avec les méthodes précédentes. En effet, on n'observe pas de dédoublement de la papille ou des vaisseaux sanguins. On distingue aussi assez nettement la zone commune aux deux images. En effet, les zones présentes sur les bords (particulièrement les bords bas et droit) présentent une délimitation franche causée par la transformation et le déplacement d'une image pour se superposer à l'autre.

Pour cet exemple, la différence absolue moyenne est la même en prenant 4 ou 10 points saillants et vaut 5.82. Ce qui signifie que le déplacement estimé est le même et que 4 paires de points ont suffi pour produire une estimation stable. Par



(a) (b)

(c) (d)

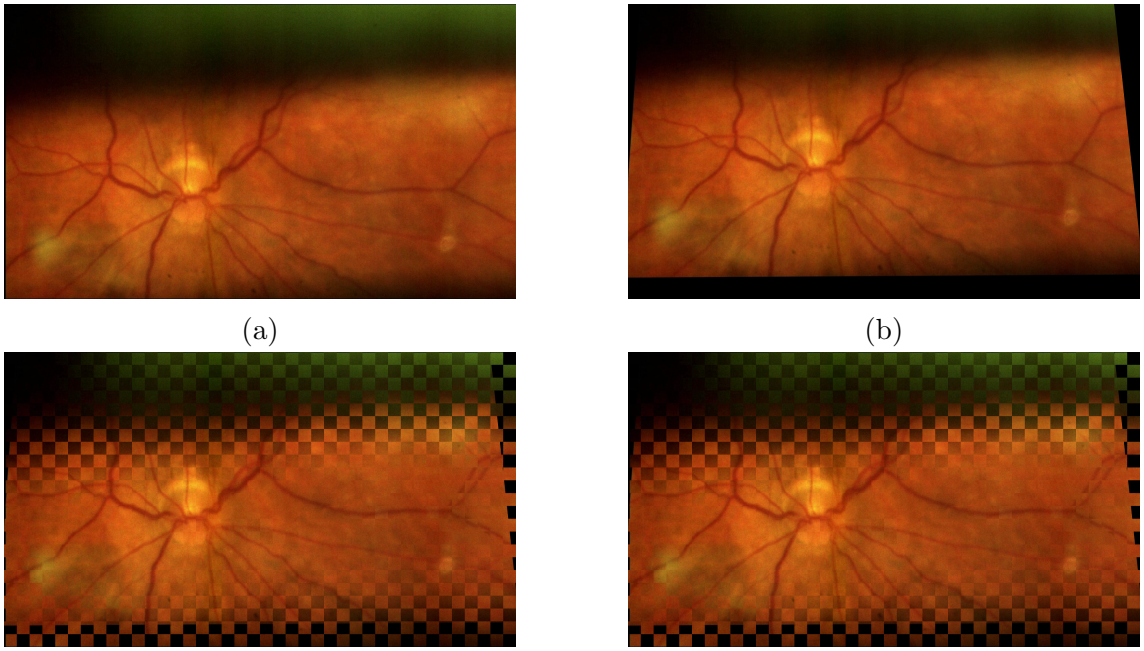
FIGURE 5.16: Exemple d'une paire d'image et des différences absolues entre la première et la seconde image (lampe à fente). a. Première image - b. Seconde image - c. Différence absolue des deux images par méthode SIFT et 4 points d'intérêt - d. Différence absolue des deux images par méthode SIFT et 10 points d'intérêt

conséquent, les damiers proposés en figure 5.17 sont eux aussi identiques pour 4 et 10 points saillants retenus par image. Visuellement, les vaisseaux sanguins ainsi que la papille sont très largement continus et dans toute la partie centrale de l'image, il est même difficile de distinguer les cases du damier. On remarque cependant un dédoublement des vaisseaux sur le bord droit de l'image. Cette déformation est potentiellement due au fait que l'image ne représente en réalité pas un plan et donc, que la courbure de l'œil n'est pas négligeable.

Sur l'ensemble des images testées, les résultats sont plus variables puisque le score global de la différence absolue moyenne vaut 53.1 dans le cas d'une estimation de déplacement avec 4 paires de points et de 18.1 dans le cas d'une estimation avec 10 paires de points. Il semble donc plus judicieux de privilégier (comme recommandé) l'option à 10 paires de points lorsque c'est possible.

SURF

Pour ce qui est de la détection de points d'intérêt par la méthode SURF, la encore suffisamment de points ont pu être détecté et mis en correspondance pour les systèmes avec 4 et 10 paires. Nous pouvons ajouter que les résultats sont d'ailleurs identiques à tous points de vue à 4 ou 10 paires. L'analyse visuelle de la différence des deux images exemple semble meilleure que celle proposée par la méthode SIFT. En effet, l'image paraît plus sombre dans la zone centrale et les contours des vaisseaux sont moins marqués, signes d'un meilleur recalage et donc d'une estimation du mouvement plus précise.



(a) (b)

(c) (d)

FIGURE 5.17: Exemple d'une paire d'image et de l'estimation des damiers composés de la première et la seconde image (lampe à fente). a. Première image - b. Seconde image recalée sur la première - c. Damier par méthode SIFT et 4 points d'intérêt - d. Damier par méthode SIFT et 10 points d'intérêt

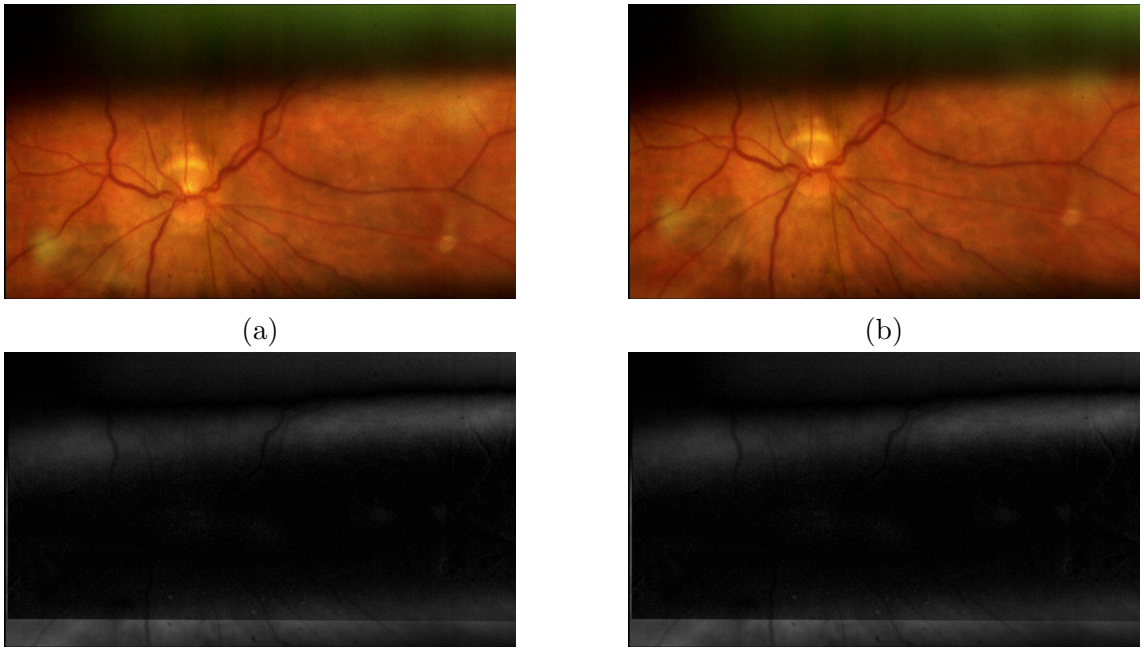
Cette impression se vérifie par les chiffres. En effet, la différence absolue moyenne pour la même zone commune que celle de la méthode SIFT vaut 5.43 (contre 5.82). Comme dans le cas précédent, le damier en figure 5.19 propose une image assez continue dans la zone centrale. On remarque également que le dédoublement des vaisseaux dans la zone droite de l'image présent dans l'exemple SIFT ne l'est pas avec SURF. Une simple homographie semble donc être en mesure d'estimer correctement le déplacement dans cet exemple. La simplification du modèle, considérant l'image comme plane, pourtant remise en question dans la partie SIFT est donc toujours valide.

De manière globale, nous observons une différence absolue moyenne de 15.1 sur l'ensemble des images testées ce qui correspond au meilleur résultat dans le tableau 5.3.

Endoscopie

En ce qui concerne les vidéos acquises à l'endoscope oculaire, nous constatons que la méthode d'estimation de mouvements basée sur la méthode SIFT n'a pas su trouver suffisamment de points d'intérêt pour permettre le calcul d'une homographie. En observant la figure 5.15 on voit que, dans notre exemple, seuls deux points d'intérêt ont pu être détecté et mis en relation.

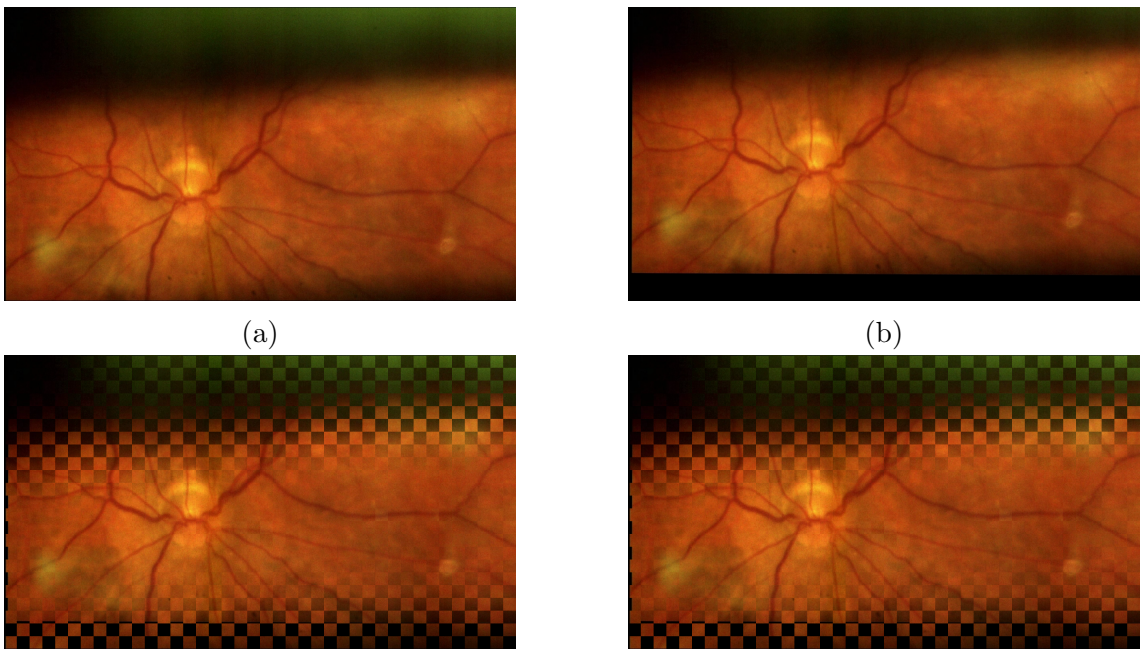
Pour la méthode SURF, nous avons obtenu des résultats pour 84 des 300 paires d'images en estimant une homographie à partir de 4 paires de points et, une fois



(a) (b)

(c) (d)

FIGURE 5.18: Exemple d'une paire d'image et des différences absolues entre la première et la seconde image (lampe à fente). a. Première image - b. Seconde image - c. Différence absolue des deux images par méthode SURF et 4 points d'intérêt - d. Différence absolue des deux images par méthode SURF et 4 points d'intérêt



(a) (b)

(c) (d)

FIGURE 5.19: Exemple d'une paire d'image et de l'estimation des damiers composés de la première et la seconde image (lampe à fente). a. Première image - b. Seconde image recalée sur la première - c. Damier par méthode SURF et 4 points d'intérêt - d. Damier par méthode SURF et 10 points d'intérêt

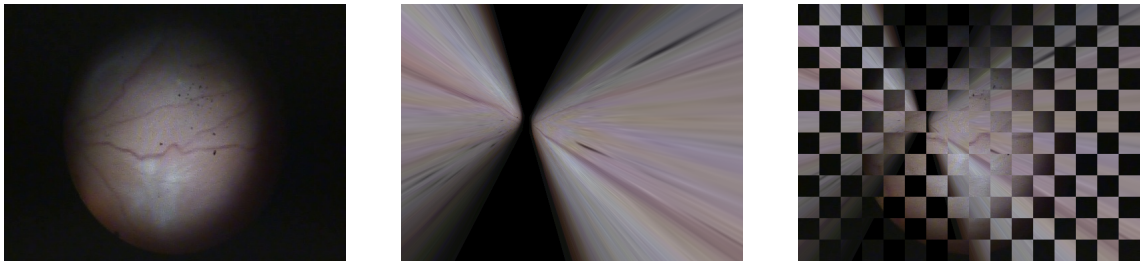
(a) (b) (c)

FIGURE 5.20: Exemple d'une paire d'image et des différences absolues entre la première et la seconde image (endoscopie). a. Première image - b. Seconde image - c. Différence absolue des deux images par méthode SURF et 4 points d'intérêt

encore, aucune paire d'images n'a permis de détecter et faire correspondre 10 paires de points. Voilà pourquoi plusieurs cases du tableau 5.15 sont vides. L'hypothèse avancée est la même que précédemment, à savoir que les images d'endoscopie (bien plus que les images de lampe à fente) sont faiblement texturées. De plus, leur faible contraste ainsi que leur plus faible définition rendent la tâche trop complexe pour de tels algorithmes.

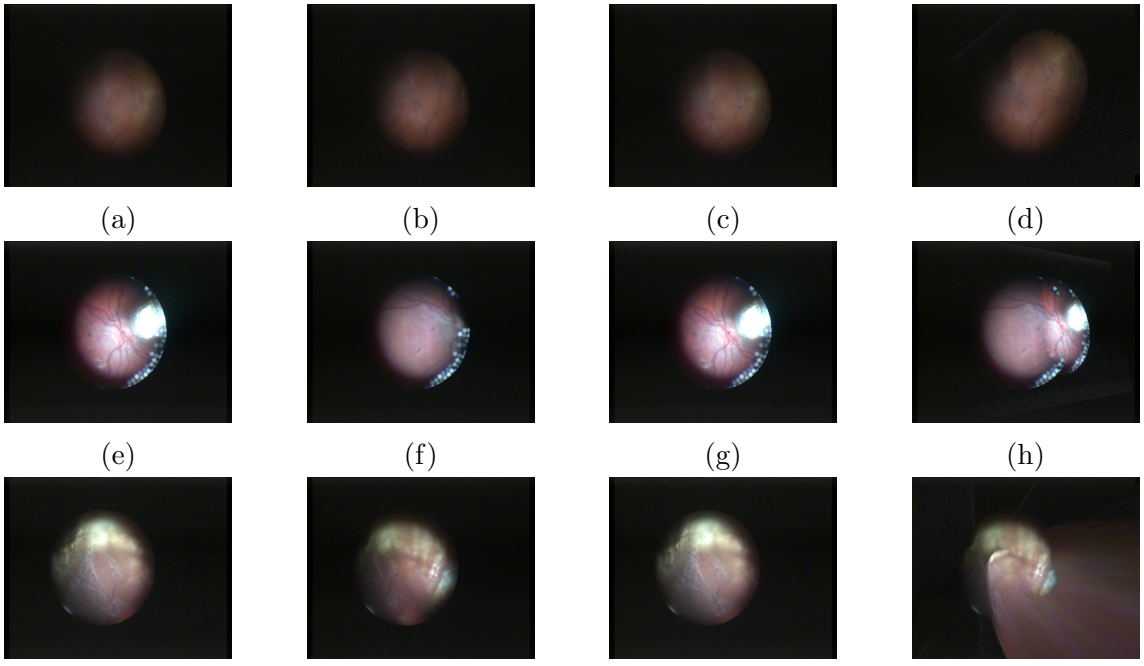
Pour le cas où 84 paires ont pu être évaluées, on constate que la différence absolue moyenne est la plus élevée du tableau 5.15, signe d'une estimation de déplacements assez fausse, voire aberrante. L'étude de l'exemple renforce cette hypothèse puisqu'il présente une différence de 146 et que visuellement l'étude de l'image différence et damier montrent un déplacement incohérent.

Afin de vérifier si l'estimation d'un mouvement de type homographique pouvait bel et bien correspondre aux mouvements des vidéos endoscopiques nous avons décidé d'annoter plusieurs paires d'images et de soumettre les correspondances manuelles de points à la méthode d'estimation de déplacement. Nous avons décidé d'annoter 6 paires d'images. Comme pour certaines paires, il était difficile de trouver précisément une dizaine de points communs. Nous en avons donc sélectionné 8. Ce nombre reste suffisant pour permettre le calcul d'une homographie. Les résultats sont visibles dans la figure 5.22 ou 3 des 6 exemples sont proposés. Les deux premières colonnes correspondent aux paires d'images étudiées. La troisième colonne montre les 8 points sélectionnés et enfin, la quatrième colonne présente la superposition de la première image et de la version recalée de la seconde. Dans deux des trois cas, on voit que le déplacement estimé semble cohérent, mais dans le troisième exemple, on retrouve un déplacement aberrant. En réalité de telles erreurs viennent du fait que pour une estimation efficace et optimale, les points sélectionnés doivent être espacés dans l'image et non pas regroupés autour d'un même axe comme c'est le cas dans le troisième exemple. Le mouvement homographique semble donc bien correspondre aux déplacements qui se produisant dans les vidéos de lampe à fente ainsi qu'en endoscopie oculaire.



(a) (b) (c)

FIGURE 5.21: Exemple d'une paire d'image et de l'estimation des damiers composés de la première et la seconde image (endoscopie). a. Première image - b. Seconde image recalée sur la première - c. Damier par méthode SURF et 4 points d'intérêt



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

FIGURE 5.22: Premier exemple de détermination manuelle de points d'intérêt. a. Première image - b. Seconde image - c. Emplacement des points d'intérêt estimés manuellement sur la première image - d. Recalage et superposition des deux images - e. Première image - f. Seconde image - g. Emplacement des points d'intérêt estimés manuellement sur la première image - h. Recalage et superposition des deux images - i. Première image - j. Seconde image - k. Emplacement des points d'intérêt estimés manuellement sur la première image - l. Recalage et superposition des deux images

5.3.2 Bilan

A l'issue des expériences réalisées on constate que la méthode SURF semble estimer assez efficacement les déplacements dans les vidéos acquises à la lampe à fente. En revanche, aucune des méthodes classiques testées ne permet d'estimation correcte pour les vidéos d'endoscopie. Nous avons vu que là où les méthodes automatiques échouent, il est aussi difficile pour un œil humain d'annoter ces vidéos. En effet, il est parfois compliqué de trouver suffisamment de points d'intérêt de type croisement de vaisseaux ou changements de texture notable sans se poser la question de leur répartition sur l'image. Le chapitre suivant cherche à résoudre ce problème d'estimation de déplacement, mais cette fois avec des méthodes utilisant l'apprentissage profond.

	Farneback	Diamond Search	SIFT 4	SIFT 10	SURF 4	SURF 10
Lampe à fente	22.1	23.0	53.1	18.1	15.1	15.1
Endoscopie	45.3	47.5	–	–	146.0*	–

TABLE 5.3: Tableau récapitulatif des erreurs absolues moyennes pour chaque méthode.
** : résultat obtenu sur 84 paires d'images et non 300*

6

Méthodes utilisant l'apprentissage profond

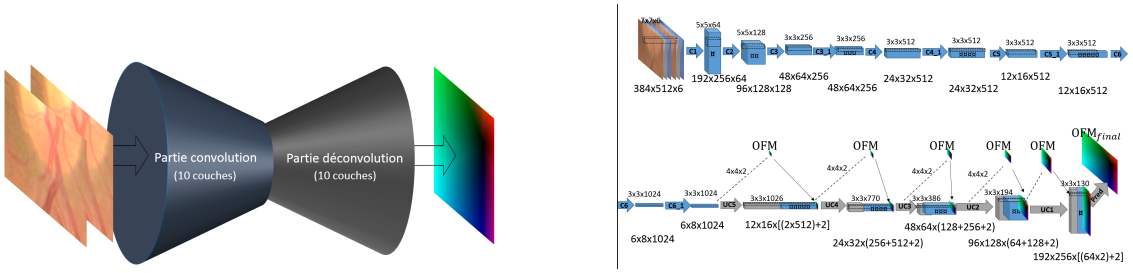
Ce chapitre se consacre à l'utilisation de méthodes utilisant l'apprentissage profond pour estimer les déplacements des vidéos acquises à la lampe à fente ou à l'endoscope oculaire. Il se divise en deux parties. La première aborde l'estimation des déplacements via les différents réseaux de neurones à convolution à apprentissage supervisé de type FlowNet. La seconde partie se consacre à un réseau à apprentissage auto-supervisé qui estime conjointement une carte de profondeurs et les déplacements à partir d'images en deux dimensions.

6.1 Estimation des déplacements via FlowNet

L'appellation FlowNet désigne un ensemble de CNN développé par des chercheurs de l'université de Fribourg en Allemagne [1],[2]. Ce sont des réseaux de neurones à apprentissage fortement supervisé spécialisés dans l'estimation de déplacements entre deux images. En effet, suite à une étape d'apprentissage, les réseaux FlowNet sont capables de fournir, en sortie, une carte de flux optique traduisant les déplacements entre les deux images proposées en entrée. Les différentes architectures de réseaux ainsi que leurs performances sur nos bases de données sont développées dans les sections suivantes. Le choix d'utiliser ce type de réseau pour notre problématique s'est fait naturellement puisqu'il s'agit de la référence dans le domaine.

6.1.1 Flownet Simple

Le premier réseau proposé est FlowNet Simple. Il apparaît dans l'article [1]. Dans cet article, on retrouve également le réseau FlowNet Corr explicité par la suite. Ces deux réseaux prennent en entrées deux images couleurs de taille 384×512 et proposent en sortie une estimation de flux optique de taille 192×256 pixels.



(a) (b)

FIGURE 6.1: Shémas de FlowNet Simple : a. simplifié - b. détaillé

Les cartes de flux optique sont plus petites que les images d'entrée, car l'équipe n'a pas constaté d'amélioration de qualité entre les cartes estimées de taille 384×512 et celles de taille 192×256. Par soucis de gain de temps elle propose donc des cartes plus petites qui peuvent être redimensionnées, à posteriori, par l'utilisateur.

Le premier réseau s'appelle FlowNet Simple, car la fusion de l'information des deux images d'entrée se fait simplement dès la première couche du réseau en empilant les images de manière à obtenir une matrice de taille $384 \times 512 \times 6$ (correspondant à la concaténation des canaux couleur des deux images cf. figure 6.1).

FlowNet Simple est un réseau encodeur-décodeur. Sa partie encodeur est composée de dix couches de convolution et sa partie décodeur de dix couches de déconvolution. En figure 6.1, un schéma global du réseau est proposé ainsi qu'un schéma plus détaillé. La partie convolution suit une structure classique de réseau de neurones encodeur. La partie déconvolution consiste à produire après chaque paire de couches, une carte de flux optique plus grande et plus précise qu'à la couche précédente. Pour ce faire, une déconvolution de la carte de caractéristiques produite par la couche précédente est effectuée. S'en suit une concaténation de celle-ci avec la carte de caractéristiques de la partie convolution ayant les mêmes dimensions et une estimation sur-échantillonnée de carte de flux optique obtenue en couche précédente. Toutes les deux déconvolutions, la résolution de la carte est augmentée d'un facteur deux.

6.1.2 Les autres Flownet

FlowNet Corr

L'architecture de FlowNet Corr (pour corrélation) est similaire à celle de FlowNet simple, d'ailleurs sa partie déconvolution est identique. La différence se fait dans la mise en commun des informations des deux images. En effet, les deux images sont introduites séparément dans le réseau et passent en parallèle dans deux branches identiques de couches de convolution. La fusion de l'information des images se fait plus tard dans le réseau, par la mise en place d'une couche de corrélation qui vient réunir les deux branches. Cette couche compare les patchs (ou cartes de caractéristiques) d'une couche avec tous les patchs de la couche correspondante de l'autre branche autour d'un certain voisinage. La corrélation entre deux patchs est établie lorsque la somme de chaque composante des patchs est minimale. Elle se fait entre les troisièmes et la quatrième couche de convolution, comme le montre la figure 6.2.

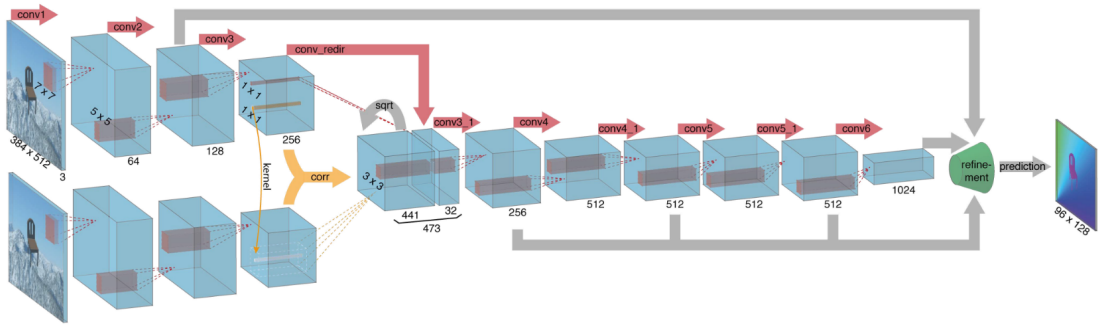


FIGURE 6.2: Schéma de détaillant la partie convolution de FlowNet Corr.
Issu de [1]

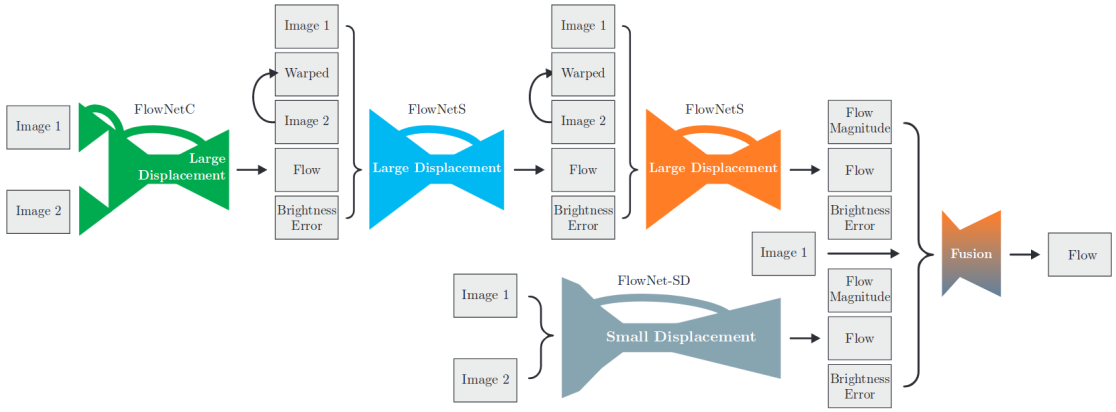


FIGURE 6.3: Schéma de la structure de FlowNet 2.
Issu de [2]

FlowNet 2

FlowNet 2 est un réseau provenant de [2]. Il est composé d'une fusion des architectures FlowNet évoquées précédemment, comme le montre la figure 6.3. Il se compose d'un enchaînement du réseau FlowNet Corr puis de deux réseaux FlowNet Simple. Une seconde branche spécialisée dans l'estimation de petits déplacements est également ajoutée. C'est une version légèrement modifiée de FlowNet Simple où la taille des noyaux de convolution et des pas sont modifiés dans les premières couches. Les deux branches sont entraînables et utilisables séparément, mais selon [2] la combinaison des deux est la solution optimale. On les trouve sous les appellations FlowNet SD pour "Small Displacements" qui est la branche dédiée à l'estimation des petits déplacements et FlowNet CSS pour "Corr-Simple-Simple" qui correspondant à l'ordre d'enchaînement des trois réseaux qui le compose.

Comme pour les autres réseaux de type FlowNet, ce réseau prend en entrée une paire d'images couleurs de taille 384×512 est propose en sortie une estimation du flux optique entre ces deux images. Les sous-réseaux intermédiaires qui le composent prennent également en entrée les deux images. Ils prennent aussi l'estimation du flux du sous réseau précédent ainsi que la seconde image recalée sur la première et enfin, la différence entre ces deux images.

6.1.3 Résultats

Les premiers entraînements/tests des réseaux FlowNet ont été réalisés sur les bases Flying Chairs et Sliding Retinas (I et II) permettant de sélectionner le CNN proposant les meilleures estimations de déplacements pour des images de rétines. Ces résultats sont présentés dans la première partie de la section. Une fois la sélection du réseau faite, d'autres tests sur les bases de vidéos acquises à l'endoscope et à la lampe à fente ont été effectués. Les résultats sont proposés dans la seconde partie.

Résultats préliminaires

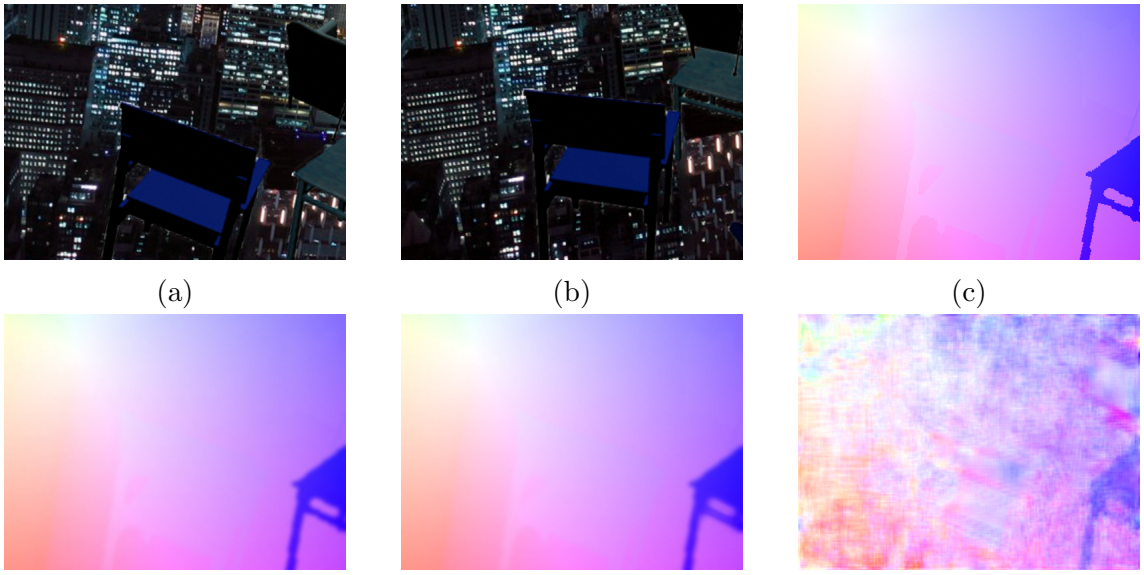
Flying Chairs

Dans un premier temps, les trois réseaux ont été entraînés sur la base Flying Chairs, qui est la base développée par les créateurs des réseaux FlowNet, pour ce type d'architecture de réseau. Les entraînements ont été faits avec les paramétrages énoncés dans les articles [1] et [2]. La seule différence est que nous faisons nos entraînements et tests sur des cartes graphiques (GPU) NVIDIA 1080 (comme dans [2]) et 1080 TI et pas des GPU NVIDIA GTX Titan comme dans [1]. Mais cette différence n'est pas susceptible d'engendrer de modification dans les apprentissages ou les tests, hormis pour les temps de calculs.

Nos résultats pour les performances de ces réseaux sont présentés dans le tableau 6.1. Les erreurs absolues moyennes sont obtenues sur 640 paires d'images non vues par le réseau pendant l'entraînement (comme dans [1]). Les trois premières lignes présentent les tests faits sur la base Flying Chairs. Les résultats que nous obtenons sont légèrement moins bons que ceux obtenus dans [1] pour FlowNet Simple et FlowNet Corr, mais restent dans le même ordre de grandeur. Ce léger écart est certainement dû aux initialisations des poids des réseaux (faites de manière automatique) qui ont été différentes entre nos entraînements et ceux réalisés par les auteurs. En ce qui concerne FlowNet 2, qui n'a été testé ni dans [1] ni dans [2] sur la base Flying Chairs, nous obtenons les moins bons résultats. Ce dernier, plus complexe et plus récents semble être le plus performant sur d'autres bases d'après [2] mais ce n'est pas le cas dans notre étude. Au vu de ces résultats et de l'image figure 6.4f, nous pouvons penser que la convergence du réseau n'a pas été atteinte, mais après avoir prolongé l'entraînement nous constatons que la fonction de coût n'évolue pas et que les flux optiques gardent le même aspect. Visuellement, les flux optiques (figures 6.4d et 6.4e) obtenus par FlowNet Simple et Corr sont très proches. Celui de FlowNet Simple parait plus précis (couleurs moins atténuées) avec des contours plus marqués.

Sliding Retinas

La base Sliding Retinas II ayant des caractéristiques très proches de la base Flying Chairs, nous avons conservé les mêmes paramétrages des réseaux que pour l'entraînement précédent. Les résultats visuels montrent qu'aucun des trois réseaux n'a été capable d'apprendre à estimer correctement les flux de la base Sliding Retinas II. Les résultats sont les mêmes pour la base Sliding Retinas I. Les lignes 4, 5 et 6 de



(a) (b) (c)

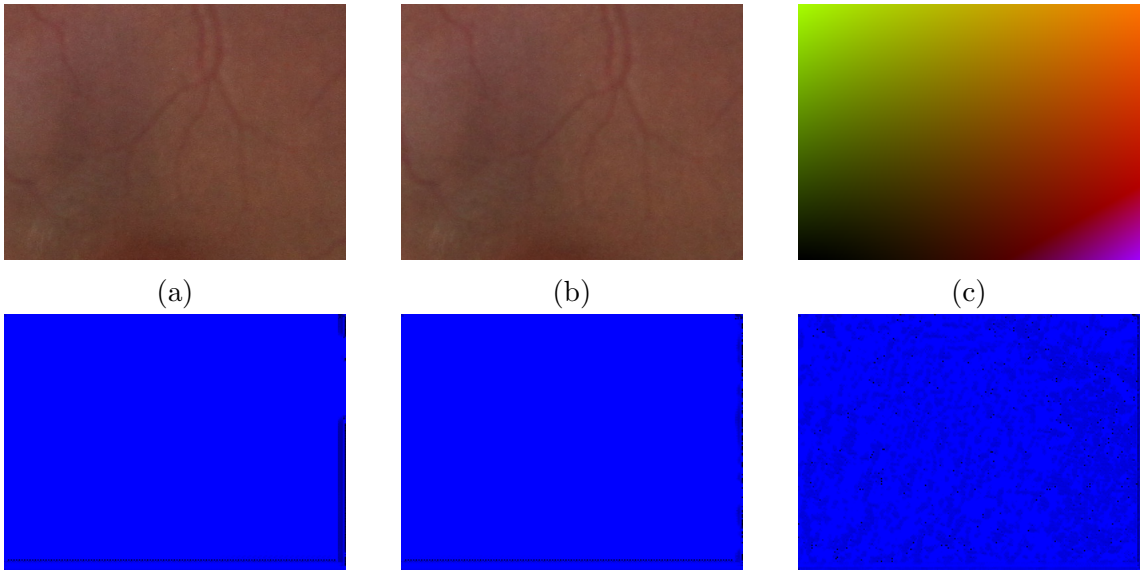
(d) (e) (f)

FIGURE 6.4: Exemples de résultats sur la base Flying Chairs.
a. Première image - b. Seconde image - c. Flux optique vérité terrain correspondant aux deux images
d. Flux optique obtenu avec le réseau FlowNet Simple - e. Flux optique obtenu avec le réseau FlowNet Corr - f. Flux optique obtenu avec le réseau FlowNet 2

la dernière colonne du tableau 6.1 confirment les résultats visuels puisqu'ils révèlent les erreurs les plus élevées. On peut cependant remarquer que même lors de l'échec de l'apprentissage on retrouve une proximité visuelle dans les résultats de FlowNet Simple et Corr et que les flux de FlowNet 2 gardent cette texture "spongieuse". On remarque également que pour les trois réseaux l'erreur est nettement plus faible en faisant des tests sur les Flying Chairs même si cette base n'a pas servi pour l'apprentissage. Nous expliquons cette différence par le fait que dans la base des Flying Chairs l'amplitude globale des déplacements est plus faible que pour les Sliding Retinas. En effet, les chaises ont des amplitudes moyennes de déplacement comparables à celles des rétines mais les fonds présents sur la base des Flying Chairs se déplacent en moyenne beaucoup moins.

Enfin, nous avons testé l'apprentissage par transfert. Les trois réseaux ont été pré-entraînés sur les Flying Chairs et affinés sur les bases Sliding Retinas. Les résultats présentés sont les meilleurs que nous avons obtenus parmi plusieurs apprentissages réalisés. En effet, nous avons fait évoluer plusieurs paramètres comme le temps de pré-apprentissage sur le premier réseau, le temps d'affinage sur le second ou encore les taux d'apprentissage. Les changements effectués sur ce dernier n'ont pas apporté de différences sur les résultats et la répartition globale de l'apprentissage optimale dans notre cas fut de consacrer 10 ères (epochs en anglais) au pré-entraînement et 10 autres pour l'affinage. Une ère correspond au passage dans le réseau de la base d'apprentissage dans son intégralité.

En ce qui concerne les tests sur la base Sliding Retinas II, visuellement les cartes de flux optiques les plus fidèles semblent être proposées par le réseau FlowNet Simple



(a) (b) (c)

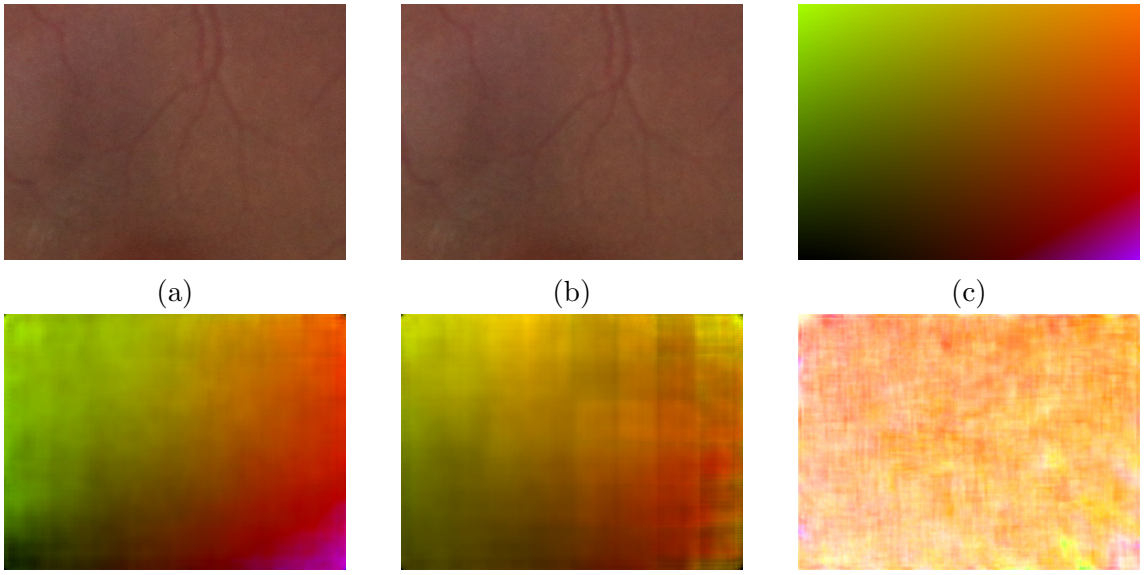
(d) (e) (f)

FIGURE 6.5: Exemples de résultats d'apprentissage direct sur la base Sliding Retinas II.
a. Première image - b. Seconde image - c. Flux optique vérité terrain correspondant aux deux images
d. Flux optique obtenu avec le réseau FlowNet Simple - e. Flux optique obtenu avec le réseau FlowNet Corr - f. Flux optique obtenu avec le réseau FlowNet 2

comme le montre la figure 6.6. Nous pouvons vérifier ce constat en analysant les trois dernières lignes de la dernière colonne du tableau 6.1. En effet, en plus d'être le meilleur résultats des trois réseaux pour la partie apprentissage par transfert, la version du réseau FlowNet Simple pré-entraîné sur Flying Chairs et affinée sur Sliding Retinas II présente l'erreur moyenne la plus faible pour l'estimation des déplacements d'images de rétines. Pour information, cette même version du réseau testée sur la base Sliding Retinas I donne une erreur moyenne de déplacement en pixel de 0.62. Là encore l'erreur moyenne est inférieure au pixel. La valeur est plus faible que pour la version II de la base sans doute parce que la version I est composée de déplacements plus simples sans changement d'échelle.

On observe bien que ce réseau s'est spécialisé, car il donne de moins bon résultats sur la base des chaises. Dans une bien moindre mesure, il en est d'ailleurs de même pour la version apprentissage par transfert du réseau FlowNet Corr. Le réseau FlowNet 2, plus récent et complexe était, sur le papier, supposé donner de meilleurs résultats que les deux précédents, mais pour l'ensemble de nos tests, c'est lui qui est le moins précis. Les déplacements qui composent nos bases d'entraînement et test ne sont peut-être pas compatible avec l'architecture du réseau. Une seconde hypothèse est que nous n'avons nous pas su le paramétrer de manière optimale malgré plusieurs tentatives et le suivi des recommandations de l'article [2]. Nous avons cependant réussi à obtenir une estimation de déplacement bien plus précise que dans les articles [1] et [2] avec le réseau FlowNet Simple.

Pour l'ensemble des raisons précédentes et surtout parce que le but final est



(a) (b) (c)

(d) (e) (f)

FIGURE 6.6: Exemples de résultats sur la base Sliding Retinas II par apprentissage par transfert.
a. Première image - b. Seconde image - c. Flux optique vérité terrain correspondant aux deux images
d. Flux optique obtenu avec le réseau FlowNet Simple - e. Flux optique obtenu avec le réseau FlowNet Corr - f. Flux optique obtenu avec le réseau FlowNet 2

Réseau	Entrainement \ Test	Flying Chairs	Sliding Retinas II
FlowNet S	Flying Chairs	**2.80**	6.25
FlowNet C	Flying Chairs	**4.57**	6.17
FlowNet 2	Flying Chairs	**7.81**	8.98
FlowNet S	Sliding Retinas II	5.20	**26.50**
FlowNet C	Sliding Retinas II	6.82	**26.61**
FlowNet 2	Sliding Retinas II	9.52	**25.83**
FlowNet S	Sliding Retinas II*	5.23	**0.69**
FlowNet C	lSliding Retinas II*	5.87	**4.40**
FlowNet 2	lSliding Retinas II*	6.31	**9.07**

TABLE 6.1: Tableau récapitulatif des erreurs absolues moyennes pour chaque méthode.
** : résultats obtenus suite à un pré-entraînement du réseau sur Flying Chairs.*
Les résultats en gras correspondent aux résultats principaux et sont illustrés par des exemples en image. Le résultat souligné correspond à l'estimation la plus précise sur la base Sliding Retinas II.

d'estimer des déplacements entre deux images de rétines, nous choisissons de faire la suite de nos tests sur la version du réseau FlowNet Simple pré-entraînée sur Flying Chairs et affinée sur Sliding Retinas II.

Résultats de FlowNet Simple sur les bases de vidéos

Pour la section suivante, le réseau utilisé suit donc les consignes du paragraphe précédent. Ne disposant pas de vérité terrain de déplacements pour les bases de vidéos d'endoscopie et de lampe à fente, comme dans le chapitre cinq, une comparaison visuelle des résultats est proposée ainsi qu'un calcul de la différence absolue entre la première image et la seconde image recalée.

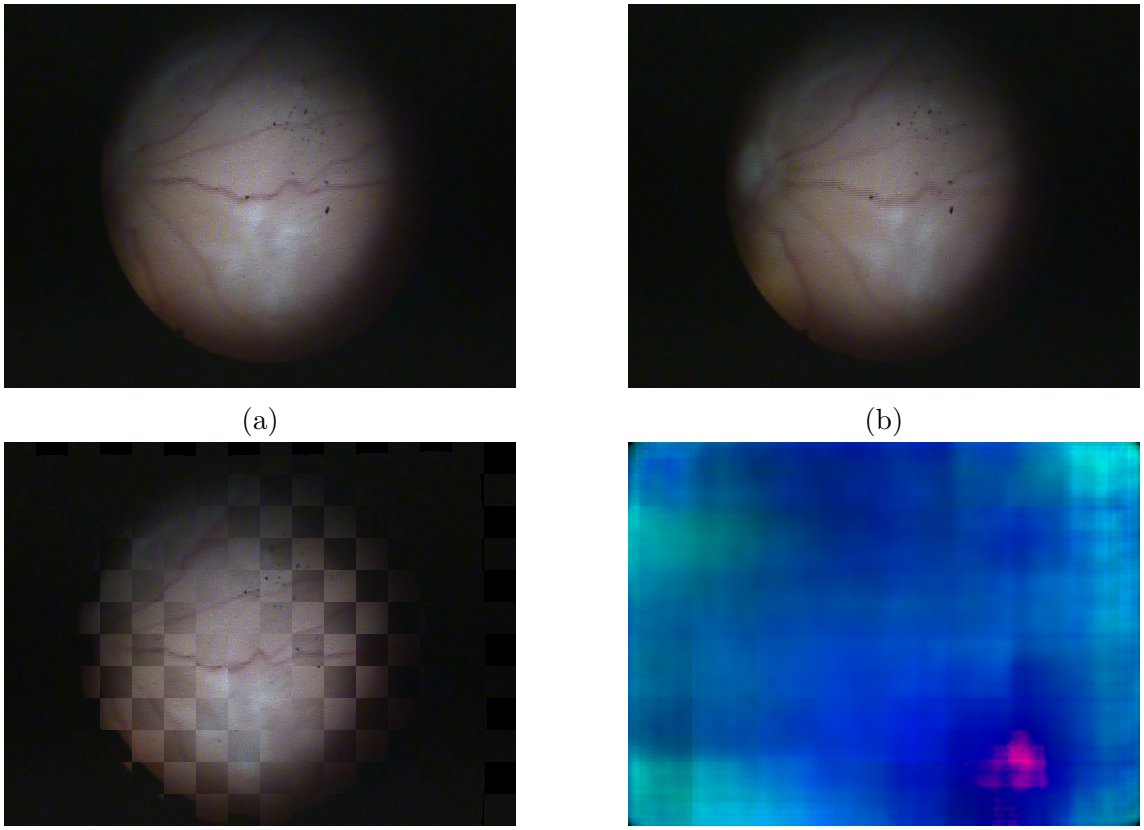
Le damier présenté en Figure 6.7c est la composition de l'image figure 6.7a et de l'image figure 6.7b ayant subi le déplacement estimé et illustré par la carte de flux figure 6.7d. On distingue des variations d'intensité de l'éclairage, mais globalement les vaisseaux semblent bien recalés. En effet, pour les vaisseaux principaux on observe qu'une certaine continuité est respectée. Pour cet exemple, l'erreur absolue moyenne est de 15.97. Cette valeur peut paraître élevée en comparaison d'un résultat visuel plutôt satisfaisant proposé par le damier. Comme précisé dans le chapitre précédent cette valeur n'est pas significative en elle-même et c'est sa comparaison avec les autres valeurs qui apporte de l'information. La raison pour laquelle la valeur reste élevée malgré un recalage plutôt bon est due à la différence d'éclairage entre les deux acquisitions. En effet, cette modalité d'évaluation n'est pas robuste aux changements d'éclairage tandis que le réseau l'est bel et bien.

Le damier présenté en Figure 6.8c est la composition de l'image figure 6.8a et de l'image figure 6.8b ayant subi le déplacement estimé et illustré par la carte de flux figure 6.8d. Pour les vaisseaux sanguins dans la partie gauche et supérieure droite du damier le recalage semble parfait. Il en est de même pour la papille. En revanche, on note quelques légères discontinuités pour des vaisseaux de la partie inférieure droite du damier. Celles-ci sont cependant assez faibles et n'entraient pas de clairs dédoublements de vaisseaux qui ont pu être observés dans les exemples du chapitre précédent. L'erreur absolue moyenne pour cet exemple est de 6.06. Dans cet exemple, les variations d'éclairages entre les deux prises de vues est nettement plus faible que pour l'exemple pris pour illustrer les résultats sur la base d'endoscopies.

Le tableau 6.2 compare les résultats obtenus avec FlowNet Simple (FNS, en gras sur le tableau) et les résultats obtenus avec les méthodes classiques présentées dans le chapitre précédent. Ce sont donc les mêmes 300 paires d'images qu'au chapitre cinq qui ont été prises pour réaliser ces tests. On distingue clairement que pour les deux bases de données la méthode basée CNN propose de meilleurs résultats. Les résultats présents sur la dernière colonne du tableau sont légèrement plus élevés que ceux des exemples vus en figure 6.7 et 6.8 mais les grandeurs restent comparables. Ces deux exemples ne sont donc pas des cas isolés où la méthode donne des résultats satisfaisants.

	Farneback	DS	SIFT 4	SIFT 10	SURF 4	SURF 10	**FNS**
Lampe à fente	22.1	23.0	53.1	18.1	15.1	15.1	**9.01**
Endoscopie	45.3	47.5	–	–	146.0*	–	**19.54**

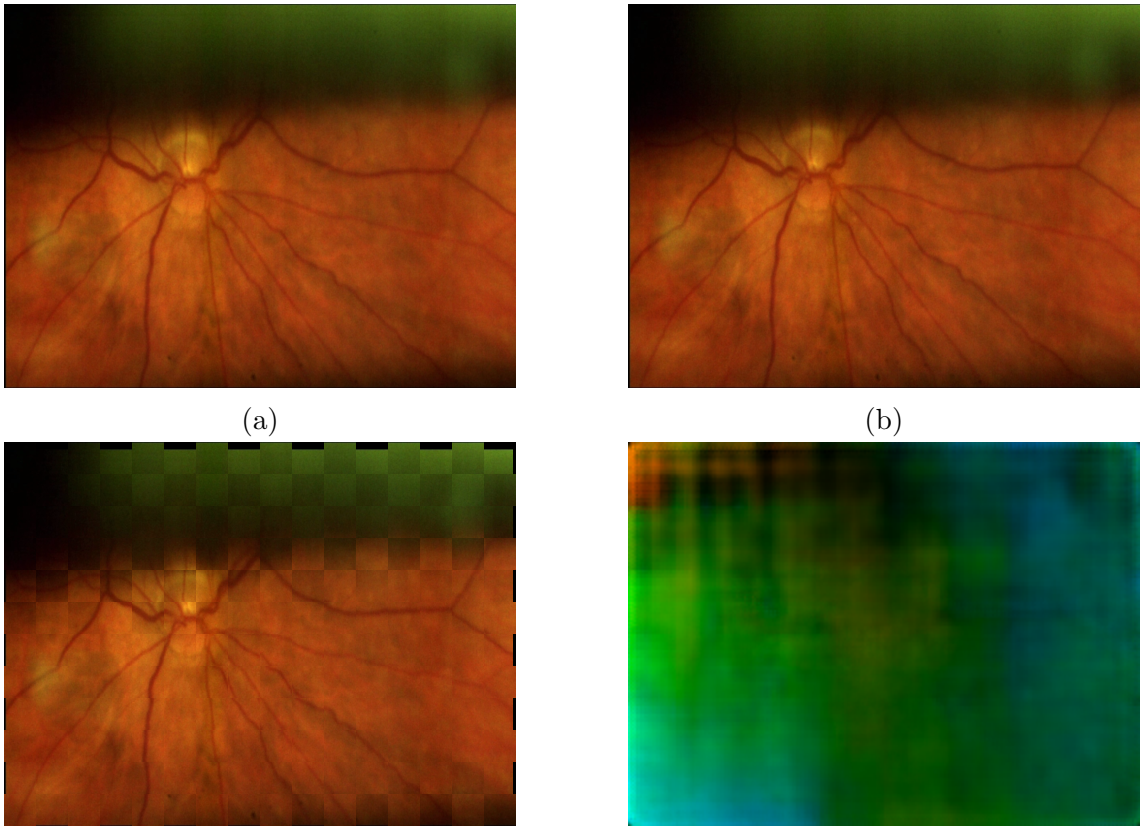
TABLE 6.2: Tableau récapitulatif des erreurs absolues moyennes pour chaque méthode.
** : résultat obtenu sur 84 paires d'images et non 300*



(a) (b)

(c) (d)

FIGURE 6.7: Exemples de résultats sur la base des vidéos acquises à l'endoscope oculaire par apprentissage par transfert du réseau FlowNet Simple.
a. Première image - b. Seconde image
c. Damier composé de a et b - d. Flux optique entre a et b obtenu avec le réseau FlowNet Simple



(a) (b)

(c) (d)

FIGURE 6.8: Exemples de résultats sur la base des vidéos acquises à la lampe à fente par apprentissage par transfert du réseau FlowNet Simple.
a. Première image - b. Seconde image
c. Damier composé de a et b - d. Flux optique entre a et b obtenu avec le réseau FlowNet Simple

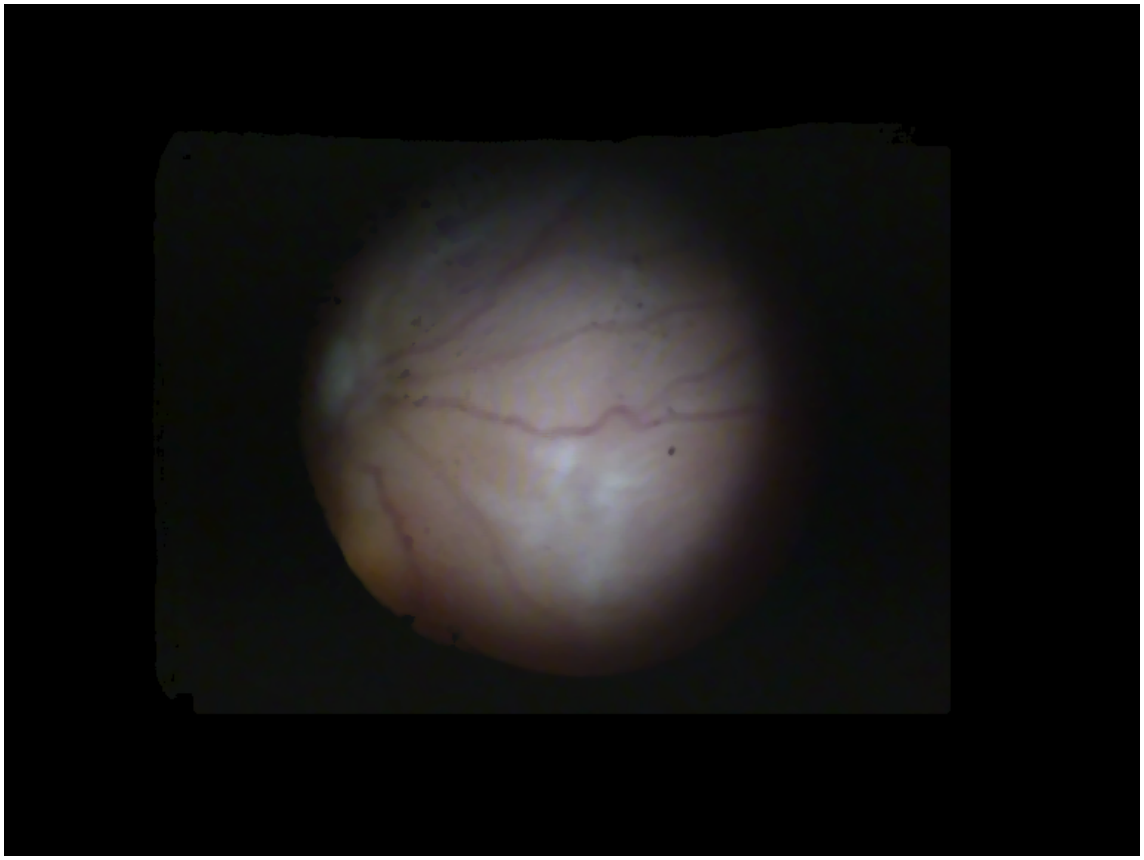


FIGURE 6.9: Mosaïque composée de 3 images de la base de vidéos acquises à l'endoscope.

6.1.4 Bilan

La meilleure méthode d'estimation de déplacements est donc celle obtenue avec le CNN FlowNet Simple pré-entraîné sur Flying Chairs et affiné sur Sliding Retinas II. Visuellement, les résultats de recalage observés sur les damiers figure 6.7c et 6.8c nous semblent acceptables, voilà pourquoi nous proposons en figure 6.9 et 6.10 des mosaïques d'images pour les bases de vidéos d'endoscopie et de lampe à fente. Il s'agit de la première étape vers l'objectif final qu'est la mise en place d'une carte dynamique progressive de la rétine.

La méthode de construction de la mosaïque est assez simple. Tout d'abord, la mosaïque est initialisée avec la première image qui se trouve au centre d'une image plus grande dont les bords sont complétés de bandes noires. La mise à jour de cette image se fait en déplaçant la seconde image du mouvement estimé. Pour les images qui suivent, le déplacement est constitué de la somme du mouvement actuel et des précédents. Cette méthode peut nettement être améliorée, mais permet tout de même de réaliser des mosaïques composées de 3 images pour la base de vidéos acquises à l'endoscope et 17 images pour les vidéos acquises à la lampe à fente. Une fois encore l'hypothèse sur la faible qualité des images de chirurgies par endoscopie oculaire semble la plus probable quant à la différence du nombre d'images composant les deux mosaïques.

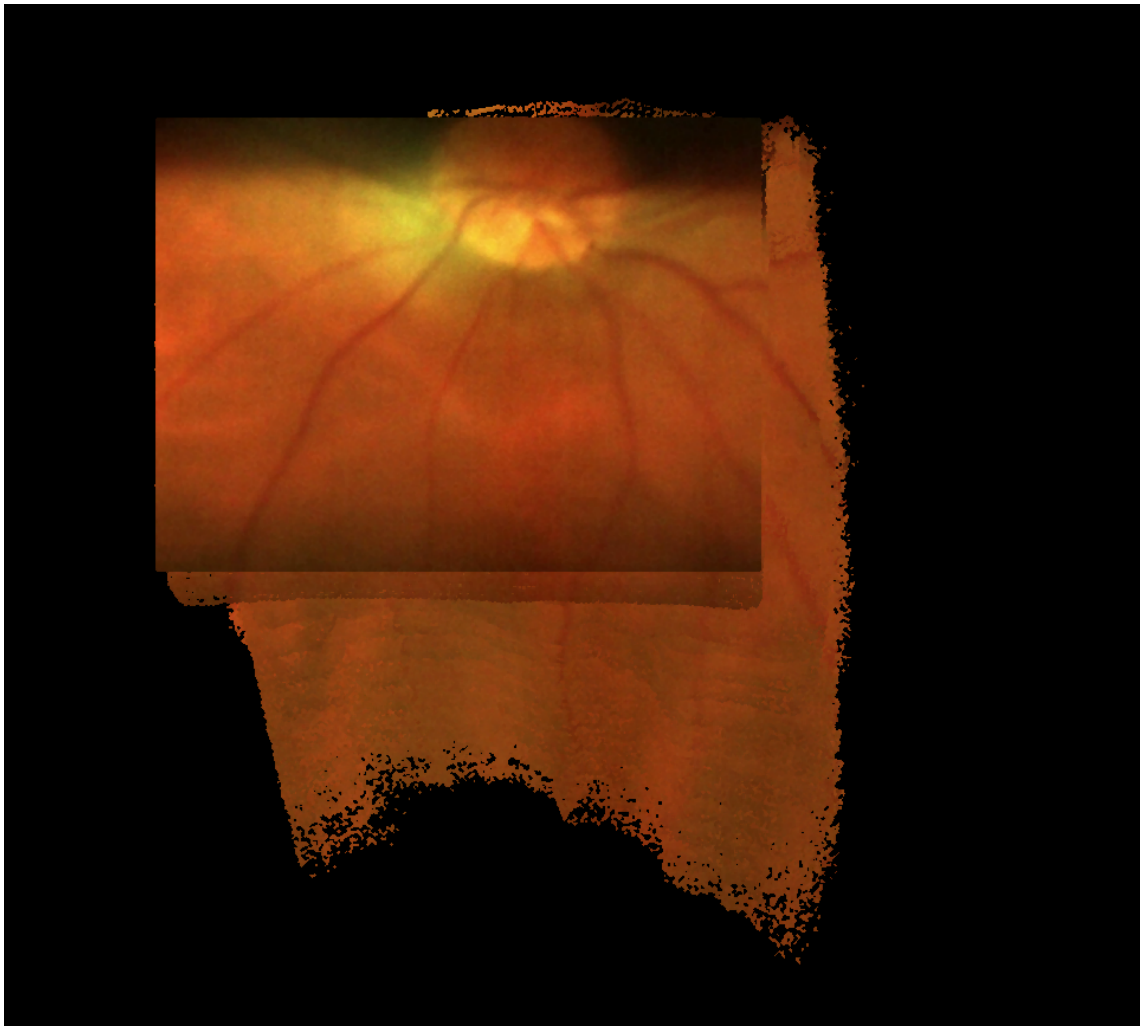


FIGURE 6.10: Mosaïque composée de 17 images de la base de vidéos acquises à la lampe à fente.

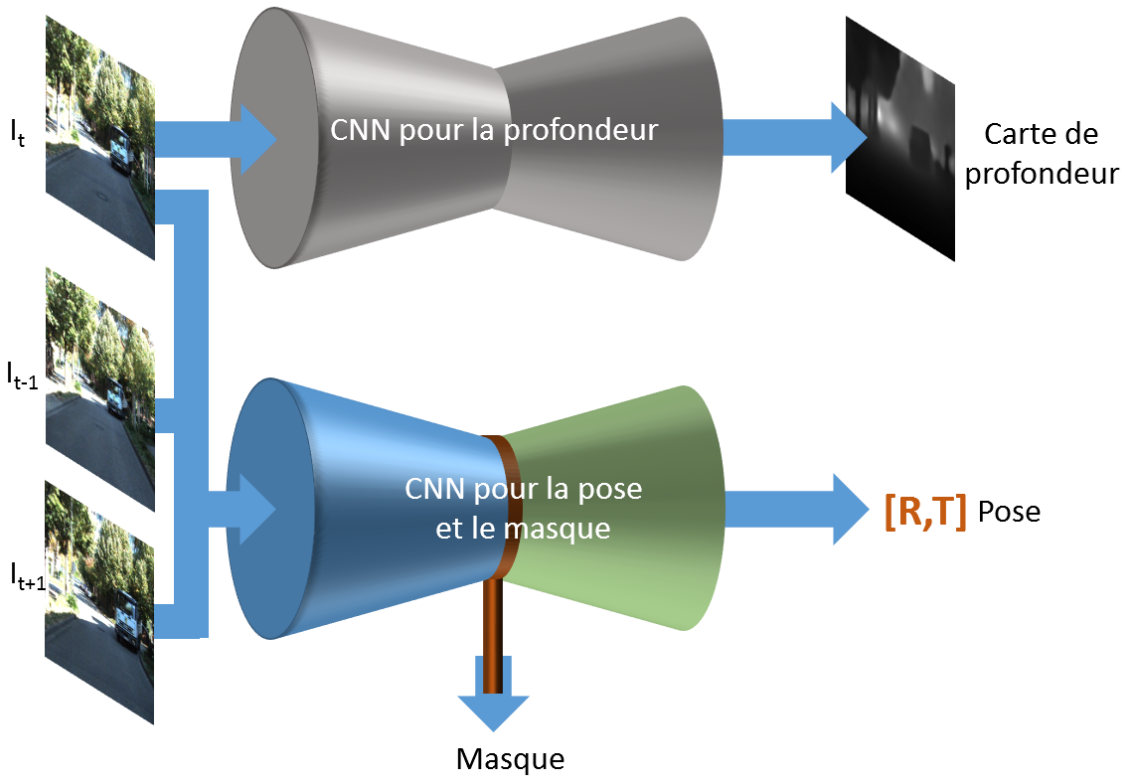


FIGURE 6.11: Schéma simplifié du réseau SFM Learner.

6.2 Estimation des cartes de profondeur

L'estimation de la profondeur d'une scène du fond d'œil peut servir à plusieurs niveaux. Cette information pourrait être utilisée pour aider au dépistage/diagnostic de certaines pathologies rétiniennes (décollement de rétine ou glaucome). Dans le cas de l'endoscopie oculaire elle pourrait aider des chirurgiens qui débutent avec ce type d'outil en les alertant si l'outil est sur le point de sortir de l'œil. Enfin, elle pourrait également servir pour la mise en place de cartes 3D de la rétine si elle est utilisée conjointement à une estimation du mouvement, comme nous allons le voir dans la suite de cette section.

Le réseau SFM Learner ([76]) est un CNN à apprentissage auto-supervisé spécialisé dans l'estimation de carte de profondeur à partir d'une prise de vue 2D. Ce réseau propose également une estimation du déplacement de la caméra entre deux prises de vues. Pour se faire, il est composé de deux réseaux principaux liés par une fonction de coût commune. Le premier CNN est de type DispNet et permet l'estimation de la profondeur. Le second calcule la pose de la caméra à partir de paires d'images fournies en entrée. Un schéma de principe du réseau est proposé en figure 6.11.

6.2.1 Méthode

Pour le CNN, de type DispNet, dédié au calcul de la profondeur, on retrouve une structure encodeur-décodeur. La partie encodeur ne prend qu'une image en entrée

et est composée de 14 couches de convolution et la partie décodeur de 7 couches de déconvolution. Comme pour la partie déconvolution des architectures FlowNet, le principe de concaténation des couches est appliqué à la partie déconvolution de ce réseau.

Le réseau dédié à l'estimation de la pose de la caméra est un simple encodeur constitué de 7 couches de convolution (5 partagées et 2 propres). Sur le même principe que la partie encodeur de FlowNetS, les couches de convolution vont extraire les caractères communs aux images fournies en entrée. La différence se fait sur la fin du réseau qui va appliquer une mise en commun générale des moyennes (global average pooling en anglais) pour rassembler toutes les prédictions des déplacements en une seule (la pose de la caméra), là où FlowNet applique un décodeur pour avoir une carte dense du flux optique. Une branche optionnelle peut être lui être ajoutée permettant la création d'un masque de confiance. Il permet de donner un poids fort aux pixels identifiés dans les deux images et un poids faible aux pixels présents dans une seule des deux images.

Sans l'activation du masque, la méthode pose comme postulat que les mouvements observés sont principalement dus au déplacement de la caméra (mouvements rigides). Elle est tout de même efficace si quelques autres mouvements sont présents. Elle considère également qu'il n'y a pas d'occlusion d'une image à l'autre. Enfin, elle suppose que toutes les surfaces sont Lambertiennes, c'est-à-dire que leur luminance est invariante en fonction de l'angle de vue. C'est notamment ce dernier point qui permet d'intégrer l'erreur photométrique à la fonction de coût du réseau. En sachant que tous ces axiomes traduisent une situation idéale, mais ne peuvent quasiment jamais s'appliquer dans des cas réels, l'ajout du masque prend tout son sens. En effet, celui-ci, va permettre de ne pas prendre en compte les zones violant une ou plusieurs des hypothèses précédentes et se replacer dans un contexte idéal. Plus tôt, si le terme de branche optionnelle a été employé pour définir ce CNN c'est parce qu'il partage sa partie encodeur avec le CNN dédié à l'estimation de la pose de la caméra. Sa partie propre est sa partie décodeur qui est composée de 5 couches de déconvolutions. La dernière couche de déconvolution va finaliser la mise en place du masque grâce à l'utilisation de fonctions softmax à sa sortie. C'est cette dernière qui permet le seuillage automatique sur la vraisemblance des zones à respecter les 3 critères précédents et leur éventuel masquage.

En plus des images, ce réseau prend en complément d'entrée les paramètres intrinsèques du dispositif d'acquisition (leur rôle est étudié dans la deuxième partie de la section résultats). Ne disposant pas de ces paramètres pour les bases d'endoscopie et d'examens à la lampe à fente, nous avons dû trouver un moyen de les retrouver. Pour se faire deux types de méthodes sont envisageables : le calibrage et l'auto-calibrage. Le calibrage est la méthode la plus simple à mettre en place et la plus efficace, mais pour la mettre en place, il faut pouvoir disposer de l'outil avant la phase d'acquisition et le faire évoluer dans un environnement spécifique et annoté.

La base de vidéos acquises par lampe à fente dont nous disposons ayant été fournie par une entreprise extérieure, nous ne pouvons intervenir en aucun cas sur le protocole d'acquisition et de pré-acquisition des images. Pour les vidéos acquises à l'endoscope, il ne nous a pas été possible d'emprunter l'outil avant intervention ou de modifier le protocole du bloc opératoire pour réaliser de courtes acquisitions pré-

opératoires. De plus, il est possible que les paramètres intrinsèques de l'endoscope évoluent si les fibres se déforment entre deux chirurgies. Nous avons donc eu recours à la seconde méthode qu'est l'auto-calibrage.

6.2.2 Auto-calibrage

L'auto-calibrage d'un dispositif d'acquisition d'images, parfois retrouvé sous l'appellation étalonnage automatique ou encore par l'anglicisme autocalibration, consiste à déterminer les paramètres internes dudit dispositif à partir d'une série acquisitions non annotées de scènes non structurées. A la différence d'un calibrage classique, l'auto-calibrage ne nécessite donc aucune mise en scène ni aucun étalonnage d'objets avant l'acquisition.

Méthode

Il permet de recréer une représentation objective du monde extérieur en compensant les effets (subjectifs) de déformations induits par le dispositif d'acquisition lui-même. Tout ceci n'est réalisable qu'en considérant l'hypothèse fondamentale selon laquelle les images sont projetées depuis un espace euclidien via une caméra de type sténopé. Il s'agit d'un dispositif composé d'un boîtier étanche à la lumière dont l'une des faces est percée d'un trou de diamètre très faible (suffisamment pour être considéré comme ponctuel). Celui-ci laisse passer la lumière vers une surface photosensible située sur la face opposée, permettant ainsi la formation d'une image renversée. Les autres faces intérieures du boîtier doivent être noires mat de manière à ne pas réfléchir les rayons lumineux. La petite taille du trou permet d'obtenir une grande profondeur de champ, parfois considérée comme infinie.

Partant de cette hypothèse, l'auto-calibrage consiste à estimer les six degrés de liberté de la caméra. A savoir la distance focale f en mm, le facteur d'échelle horizontal k_u et vertical k_v en m^{-1}, le centre de l'image ($[u_0 ; v_0]$) en px et l'inclinaison θ en radian ou degrés. Le facteur d'échelle correspond à l'inverse de la taille d'un pixel dans l'espace des coordonnées du monde. La focale f est la distance entre un plan principal et son foyer correspondant. Le centre de l'image se situe à l'intersection de l'axe optique principal et du plan principal image. Enfin, l'inclinaison est un paramètre introduit pour permettre de prendre en compte d'éventuelles déformations de la grille de pixel. Il s'agit d'un angle dont la valeur est très proche $\pi/2$. Une fois ces paramètres déterminés il est donc possible d'estimer des transformations euclidiennes à 6 degrés de liberté à partir de séquences d'images qui ne sont pas nécessairement calibrées.

En 1992 Faugeras et al. présentent une théorie mathématique pour l'auto-calibrage de caméras à partir de vues multiples ([121]). Dans celle-ci, ils démontrent qu'il faut au minimum trois vues différentes d'une même scène pour réaliser un étalonnage complet avec des paramètres intrinsèques constants, sans contrainte posée à priori sur la scène ou sur le dispositif d'acquisition. Ils précisent également qu'en se modernisant, les capteurs et les optiques de bonne qualité permettent de simplifier le problème en considérant la grille de pixels comme étant orthogonale et les pixels carrés. Ainsi, on peut passer le nombre d'images minimum de trois à deux. Enfin,

un auto-calibrage à partir d'une seule image peut être effectué en fournissant des informations sur la structure de la scène comme la taille et les distances entre certains objets qui la compose.

Dans le cas le plus global, la relation entre l'espace de coordonnées de l'image et celui du monde réel peut se traduire par l'équation mathématique (6.1),

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = A \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} G \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{6.1}$$

où $[su\,;sv\,;s]$ sont les coordonnées de l'espace image et s un facteur d'échelle différent de zéro. $[X\,;Y\,;Z]$ traduisent l'espace de coordonnées du monde réel. A est la matrice de taille 3×3 des paramètres intrinsèques de la caméra (au nombre de 6) et G est la matrice de taille 4×4 de déplacement traduisant la position et l'orientation de la caméra caractérisés par 6 paramètres (parfois appelés paramètres extrinsèques).

La méthode proposée par Faugeras et al. se décompose en deux étapes principales. La première consiste à trouver les épipoles par la méthode de Sturm [122] ou par la méthode des matrices fondamentales. Pour cette étude, nous avons choisi la méthode des matrices fondamentales car celle-ci est plus facile à développer.

La géométrie épipolaire est un modèle mathématique servant à décrire les relations entre deux prises de vues d'un même objet ou d'une même scène. Les principes mathématiques sont posés dès la fin du XIXème siècle par Hauck dans [123] mais le modèle prend toute son importance quelques dizaines d'années plus tard avec l'avènement des images numériques.

Dans ce modèle mathématique, la matrice fondamentale (souvent désignée par la lettre F) est porteuse de l'information sur la transformation qui lie les deux images. C'est une matrice de taille 3×3.

La seconde étape consiste à résoudre le système des équations de Kruppa (6.2). Elles permettent de lier la transformation, estimée grâce aux épipoles, à l'image de la conique absolue et ainsi d'obtenir les paramètres intrinsèques du système d'acquisition.

$$\frac{A_{11}}{A_{12}} = \frac{A'_{11}}{\rho A'_{12}} \qquad \frac{A_{22}}{A_{12}} = \frac{\rho A'_{22}}{A'_{12}} \tag{6.2}$$

Avec :

$$\begin{aligned} A_{11} &= -\delta_{13}p_3^2 - \delta_{12}p_2^2 - 2\delta_1 p_2 p_3 \\ A_{12} &= -\delta_{12}p_1 p_2 - \delta_3 p_3^2 + \delta_2 p_2 p_3 + \delta_1 p_1 p_3 \\ A_{22} &= -\delta_{23}p_3^2 - \delta_{12}p_1^2 - 2\delta_2 p_1 p_3 \end{aligned} \tag{6.3}$$

Où p sont les épipoles et δ sont les composants de D, le dual de la conique absolue, tels que :

$$D = \begin{bmatrix} -\delta_{23} & \delta_3 & \delta_2 \\ \delta_3 & -\delta_{13} & \delta_1 \\ \delta_2 & \delta_1 & -\delta_{12} \end{bmatrix} \tag{6.4}$$

Le concept de dualité en géométrie projective à été introduit par le mathématicien Jean-Victor Poncelet au XIXème siècle et permet de transformer des séries de points en droites/courbes (et réciproquement). Pour plus de détails sur le vocabulaire, les principes mathématiques et la résolution des équations, se reporter aux articles fondateurs [121], [124] et plus récemment [125] et [126].

Résolution des systèmes

La résolution des équations de Kruppa pour l'estimation des paramètres intrinsèques s'est faite par une méthode développée dans le cadre de la thèse. En effet, il n'existe pas, à notre connaissance, de méthode disponible en ligne permettant de retrouver les 6 paramètres intrinsèques d'une caméra.

Dans un premier temps, nous sélectionnons automatiquement 25 points d'intérêt sur huit triplettes d'images. Les points d'intérêt sont globalement placés au centre de l'image et sont séparés de proche en proche de 10 pixels. Un plus grand nombre d'images et de points d'intérêt que le minimum permet aux systèmes d'être surdéterminés et d'éliminer automatiquement les potentielles valeurs aberrantes qui peuvent se glisser les points d'intérêt.

Une fois les points sélectionnés, nous considérons les systèmes comme un problème d'optimisation à résoudre. Nous obtenons donc une série de solutions minimisant les erreurs au sens des moindres carrés. Pour chacune des vidéos, plusieurs dizaines de milliers de solutions sont trouvées. En enlevant les valeurs qui ont un sens mathématiquement, mais qui physiquement sont aberrantes (comme par exemple des valeurs de focales ou de coordonnées de centre d'image négatives) nous réduisons ce nombre à quelques centaines. Enfin, en sélectionnant les couples de solutions qui minimisent l'erreur tout en restant cohérentes avec les valeurs de centre d'image que nous estimons par mesure manuelle nous arrivons à une solution pour chacune des modalités d'acquisition. Les estimations de ces valeurs sont proposées dans le tableau 6.3. Nous n'avons pas de moyen pour évaluer précisément la véracité de ces paramètres en revanche, nous avons testé l'entraînement du réseau SFM Learner sur la base KITTI avec les bons paramètres intrinsèques et avec de faux paramètres. Ces résultats sont proposés dans la deuxième partie de la section résultats.

En ajoutant à la base de données de vidéos acquises à l'endoscope et à la lampe à fente les valeurs estimées des paramètres intrinsèques, il est désormais possible de réaliser des entraînements du réseau SFM Learner. Les résultats de ces entraînements et les différents tests sont font l'objet de la partie suivante.

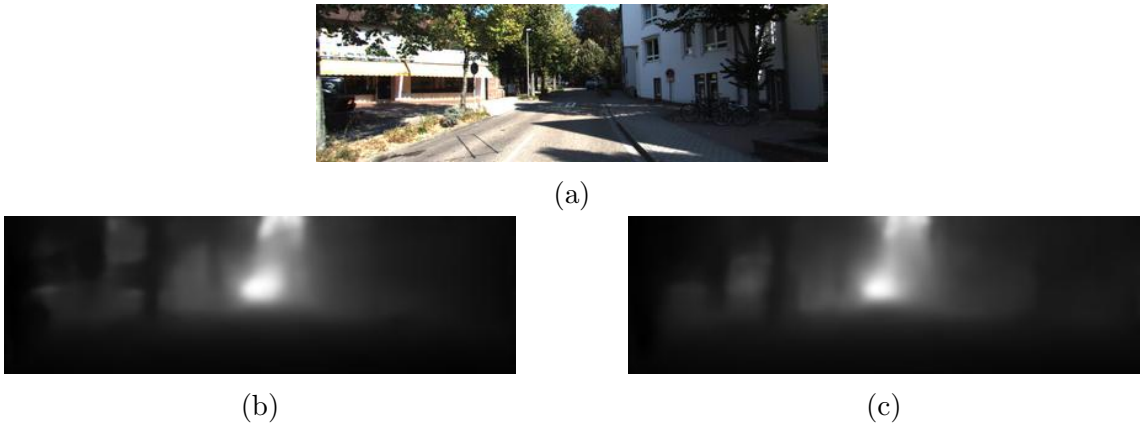
6.2.3 Résultats

Plusieurs entraînements ont été effectués. Dans un premier temps des entraînements du réseau sur la base KITTI ont été effectués dans les mêmes conditions que dans l'article [76]. Dans un seconde temps, nous avons étudié le rôle des paramètres intrinsèques sur la base KITTI. Dans un troisième temps, les images de la base KITTI

Base	f	k_u	k_v	u_0	v_0	θ
Lampe à fente	102.9	1.0	1.2	195.2	259.0	90.0
Lampe à fente	107.0	1.0	1.2	193.1	258.7	90.0
Lampe à fente	105.6	1.0	1.1	193.4	255.2	90.0
Endoscope	17.0	1.0	1.1	272.4	171.7	90.0
Endoscope	24.1	1.0	1.0	234.1	189.5	90.1
Endoscope	22.4	1.0	1.0	214.7	219.2	90.0
Endoscope	14.9	1.1	1.2	272.9	174.4	90.8
Endoscope	25.3	1.1	1.0	243.3	231.5	90.0
Endoscope	25.5	1.0	1.1	269.4	207.1	90.0
Endoscope	15.7	1.0	1.0	285.0	189.2	90.0
Endoscope	19.1	1.1	1.0	237.1	222.1	90.0
Endoscope	16.3	1.1	1.0	243.5	216.3	90.0
Endoscope	31.6	1.0	1.0	243.8	192.1	90.0
Endoscope	23.8	1.0	1.1	237.1	192.6	90.0
Endoscope	26.8	1.0	1.2	221.7	195.2	90.1
Endoscope	27.0	1.0	1.0	291.3	183.4	90.0
Endoscope	24.5	1.0	1.0	314.7	165.9	90.0
Endoscope	30.6	1.0	1.0	237.7	174.1	90.0
Endoscope	21.2	1.1	1.0	278.6	159.8	90.0
Endoscope	25.1	1.1	1.0	275.5	204.6	90.0
Endoscope	34.2	1.0	1.0	237.1	201.9	90.0
Endoscope	28.4	1.0	1.0	291.0	192.1	90.0
Endoscope	17.7	1.1	1.0	221.1	192.3	90.0
Endoscope	19.3	1.0	1.1	262.8	186.4	90.0
Endoscope	39.3	1.0	1.0	246.3	180.0	90.0
Endoscope	21.2	1.0	1.0	256.1	213.8	90.0

TABLE 6.3: Tableau récapitulatif des paramètres intrinsèques estimés par notre méthode pour les différentes vidéos de nos bases de données. Avec f en mm, k_u et k_v en m^{-1}, $[u_0 ; v_0]$ en px et θ en degrés.

ont été redimensionnées de manière à correspondre aux dimensions de nos bases de vidéos. Cette étape a pour objectif de servir de pré-entraînement au réseau et de le raffiner sur nos bases de vidéos de rétines. Enfin, des entraînements directs du réseau ont été faits avec les bases de vidéos acquises à l'endoscope et à la lampe à fente. Comme les variations d'éclairage sont bien plus fortes sur nos bases de vidéos rétiniennes que dans la base KITTI et que, de ce fait, la branche masque cacherait une grande partie des images, nous avons choisi de faire également des entraînements avec des versions de bases de données dont les intensités des images sont normalisées. Ce faisant, il est plus facile pour le réseau de faire correspondre deux zones identiques malgré des conditions d'éclairage différentes à l'acquisition. En effet, le principe de la méthode se base autour de l'erreur photométrique, minimiser les variations d'éclairage entre deux prises de vue est donc très important pour un fonctionnement optimal.



(a) (b) (c)

FIGURE 6.12: Exemples de carte de profondeurs issus de la base KITTI : a. Image originale - b. Carte estimée suite à un entraînement sans branche masque - c. Carte estimée suite à un entraînement avec branche masque

Résultats des entraînements avec et sans masque sur la base KITTI

Comme à chaque nouvel essai de réseau, nous testons dans un premier temps ses performances telles que décrites dans l'article afin de comparer nos résultats et ceux des auteurs. Nous entraînons et testons donc le réseau sur la base KITTI avec le paramétrage décrits dans [76]. Nous profitons de cette phase préliminaire pour entraîner et tester une version du réseau sans la branche masque.

Nous comparons visuellement l'estimation de la profondeur avec les exemples figure 6.12 et nous remarquons que les résultats sont assez proches. Plus la carte de profondeur est claire, plus la profondeur estimée est grande. Aussi, sur certains points, l'image issue de l'entraînement sans le masque, les contours semblent légèrement plus marqués. Mais on remarque également que sur la gauche de l'image la zone sous le haut-vent est détectée de manière assez précise mais comme étant devant ce même haut-vent. En effet, celle-ci est plus sombre que le reste du bâtiment 6.12b. Tandis que sur la figure 6.12c cette même zone est plus floue, mais est détectée comme étant derrière le haut-vent (zone plus claire que le bâtiment).

Cette remarque est appuyée par les résultats chiffrés. Ils correspondent à la comparaison des résultats estimés avec la vérité terrain de la base KITTI. Pour rappel celle-ci n'est utilisée que dans le cadre de test et l'entraînement se fait de manière auto-supervisée. Pour les 1 591 images (appelée séquence 09 dans [76] et dans KITTI) de la base de test nous obtenons une erreur moyenne absolue de 0.329 m pour la profondeur et de 0.0129 m pour l'erreur de l'estimation de la trajectoire pour la version du réseau avec le masque activé. Nous obtenons 0.381 m comme erreur pour la profondeur et 0.0206 m pour la pose pour la version avec le masque désactivé. L'écart est faible, mais comme dans le cadre de l'article, il semblerait bel et bien que cette branche "masque" joue un rôle dans le cadre de l'entraînement du réseau. Pour la suite des entraînements et tests nous avons donc gardée active la branche "masque" optionnelle.

Résultats des entraînements avec changements des paramètres intrinsèques

Comme nous ne disposons pas des paramètres intrinsèques pour nos bases de vidéos et que nous n'avons aucun moyen pour vérifier ceux que nous avons estimés, nous réalisons deux entraînements du réseau sur la base KIITI en modifiant ces paramètres. Ces tests ont pour but d'étudier l'importance que ces paramètres occupent dans la phase d'apprentissage du réseau. Dans un premier temps, nous modifions les paramètres intrinsèques de la base KITTI d'un facteur 10 et dans un second temps d'un facteur 100.

Les résultats proposés dans le tableau 6.4 sont des moyennes obtenues sur les 1 591 images de la séquence 09 de la base KITTI. Sans surprise, les meilleurs résultats sont obtenus avec les bons paramètres intrinsèques. On constate que les modifier d'un facteur 10 augmente légèrement l'erreur sur l'estimation de la profondeur, mais augmente fortement (quasiment d'un facteur 60) l'erreur sur l'estimation de la trajectoire. Naturellement, en faussant encore plus les paramètres intrinsèques, en les multipliant par 100, nous obtenons des erreurs encore plus grandes sans pour autant faire chuter drastiquement. En effet, les erreurs restent du même ordre de grandeur pour des valeurs de paramètres intrinsèques multipliées par 10 ou 100.

Erreur moyenne absolue	pour la profondeur (en m)	pour la trajectoire (en m)
Bons paramètres intrinsèques	**0.329**	**0.0129**
Paramètres $\times 10$	0.488	0.6036
Paramètres $\times 100$	0.639	0.9104

TABLE 6.4: Erreurs d'estimation de la profondeur et de la pose pour des entraînements faits avec différentes valeurs de paramètres intrinsèques (meilleurs résultats en gras).

La visualisation de l'exemple proposé en figure 6.13 confirme les résultats proposés par le tableau 6.4. On observe que l'estimation de la profondeur est de moins en moins bonne à mesure que l'on augmente l'erreur sur les paramètres intrinsèques pendant l'apprentissage. On observe également que malgré cela, le réseau arrive à apprendre quelque chose et propose des estimations non nulles.

Ces paramètres jouent donc bel et bien un rôle dans l'entraînement, mais nous voyons qu'en introduire de faux baisse les performances du réseau sans pour autant être un verrou pour son apprentissage. Nous voyons donc que, dans l'hypothèse où les paramètres que nous avons estimés pour les vidéos de rétine se seraient pas précis ou partiellement faux, l'apprentissage du réseau resterait tout de même être envisageable.

Résultats des entraînements avec masque sur KITTI redimensionné

A l'origine la taille des images à l'entrée du réseau est de 128×416 pixels or la taille de nos images est de 384×512 pixels. Nous avons donc relancé un entraînement du réseau avec un paramétrage compatible avec de telles tailles d'images. Afin de pouvoir quantifier les causes de telles modifications nous avons réalisé un premier entraînement sur la base KITTI préalablement redimensionnée. Globalement les

(a) (b)

(c) (d)

FIGURE 6.13: Exemples de carte de profondeurs issus de la base KITTI avec paramètres intrinsèques modifiés : a. Image originale - b. Carte estimée suite à un entraînement avec les bons paramètres intrinsèques - c. Carte estimée suite à un entraînement avec des paramètres intrinsèques 10 fois plus grands - d. Carte estimée suite à un entraînement avec des paramètres intrinsèques 100 fois plus grands

images redimensionnées 384 × 512 pixels sont plus grandes, mais présentent un champ plus restreint autour du point de fuite que les images 128 × 416 (plus étirées).

Malgré une taille d'images plus grande, les résultats visuels sur l'estimation de la profondeur ne semblent pas perdre en qualité et restent cohérents. Les résultats quantitatifs sur l'erreur de la profondeur et de la pose confirment le visuel. En effet, on trouve en moyenne une erreur absolue de 0.331 m pour l'estimation de la profondeur (contre 0.329 m avec les tailles originales) et 0.0190 m pour l'erreur sur la pose (contre 0.0129 m). Comme précédemment, les résultats sur la pose sont globalement meilleurs que ceux obtenus dans [76], (0.021) mais sont légèrement moins bons pour l'estimation de la profondeur (0.208 m).

Cette étape nous a permis de valider le fait que le réseau puisse être performant sur des images plus grandes que celles décrites dans [76]. Cette version du réseau nous a également servi de pré-entraînement lors de la phase d'apprentissage par transfert présentée par la suite. Nous pouvons donc passer aux essais sur les données rétiniennes.

Résultats des entraînements directs testés sur les vidéos acquises à l'endoscope et à la lampe à fente

Nous commençons par un entraînement direct du réseau sur les bases de vidéos acquises à l'endoscope et à la lampe à fente. Plusieurs tentatives de paramétrage (modification sur le taux d'apprentissage entre $1e^{-4}$ et $5e^{-8}$) du réseau donnent les mêmes résultats et, pour les deux bases, le réseau n'est pas capable d'apprendre à estimer directement la profondeur et la pose pour ce type de vidéos. Nous pouvons l'observer à travers les exemples figures 6.15c et 6.15d. Il n'est cependant pas possible de quantifier les erreurs, car contrairement à la base KITTI, nous ne disposons pas de vérité terrain pour ces bases. Nous évaluons donc visuellement la qualité de construction des cartes de profondeurs.

Dans un second temps, nous testons ces mêmes images sur la version du réseau

(a) (b)

(c) (d)

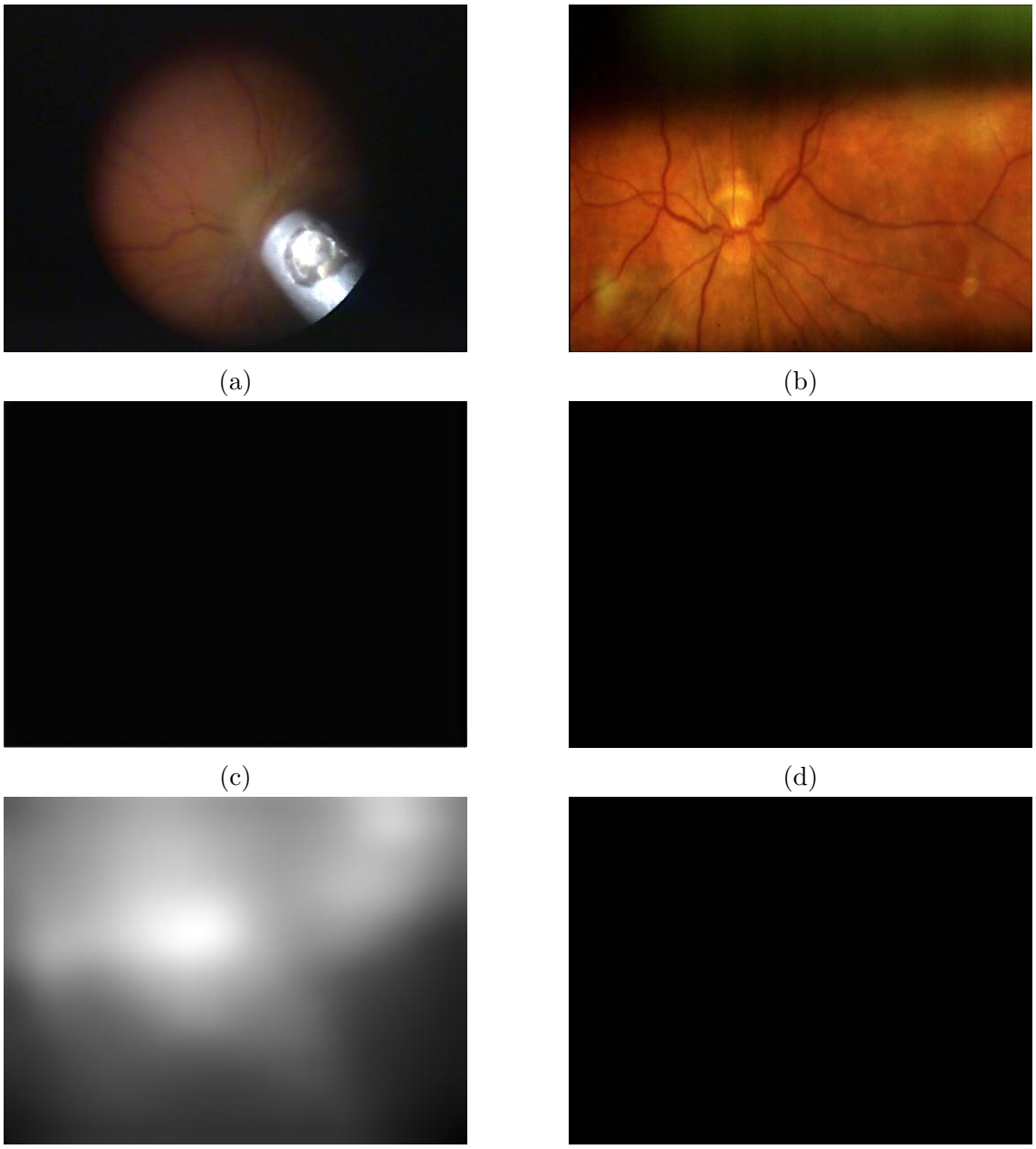
FIGURE 6.14: Comparaisons des cartes de profondeurs entre des entraînements faits à différentes résolutions : a. Image originale (128×416px) - b. Carte de profondeur estimée (128×416px) - c. Image originale (384×512px) - d. Carte de profondeur estimée (384×512px)

précédente (entraînée sur KITTI 384×512). Les résultats sont présentés en figures 6.15e et 6.15f. Nous observons que l'estimation de la profondeur ne rend pas un résultat nul pour la base acquise à l'endoscope et propose des variations de profondeurs cohérentes avec la réalité par endroits. Par exemple dans le cas présenté en figure 6.15e, le centre de l'image est plus clair (donc plus loin), une zone plus sombre se forme au niveau de l'outil. En revanche, les bords de l'image, non-porteurs d'information, sont pourtant détectés comme présentant des variations de profondeur. Il s'agit d'une incohérence majeure de l'estimateur que nous espérons corriger avec l'étape d'apprentissage par transfert. En effet, dans la base KITTI toute l'image est porteuse d'information tandis que dans cette base, seule l'ellipse centrale est considérée comme utile à l'étude. Nous espérons donc que la phase d'affinage permette au réseau de le détecter et soit incluse dans le masque.

Pour les vidéos acquises à la lampe à fente, toute l'image est porteuse d'information, mais le réseau entraîné sur KITTI ne semble pas non plus permettre une estimation non nulle de la profondeur 6.15f. Nous espérons tout de même que cette étape puisse initialiser le réseau pour la phase d'apprentissage par transfert décrite dans la partie suivante.

Résultats des apprentissages par transfert et normalisation

Encore une fois, pour cette étape, plusieurs entraînements ont été réalisés. La base du réseau est restée la même, à savoir la version du réseau SFM Learner préalablement entraînée sur KITTI (384×512px), mais l'affinage à changé. Nous avons respectivement testés des affinages avec les bases acquises à l'endoscope et



(a) (b) (c) (d) (e) (f)

FIGURE 6.15: Comparaisons des cartes de profondeurs sur des images issues de nos bases de données pour différents entraînements. a. Image acquise à l'endoscope - b. Image acquise à la lampe à fente - c. Carte de profondeur estimée après entraînement direct sur la base de vidéos endoscopiques - d. Carte de profondeur estimée après entraînement direct sur la base de vidéos acquises à la lampe à fente - e. Carte de profondeur estimée après entraînement direct sur la base KITTI (384×512px) - f. Carte de profondeur estimée après entraînement direct sur la base KITTI (384×512px)

à la lampe à fente. De plus, nous avons réalisés des entraînements avec des taux d'apprentissages variables (entre $2e^{-5}$ et $5e^{-8}$). Pour les exemples présentés dans la figure 6.16 le taux d'apprentissage variait entre $5e^{-6}$ et $5e^{-8}$ au cours de l'affinage. Ce paramétrage semble être le plus optimal que nous ayons trouvé, mais donne toutefois des résultats médiocres.

En effet comme le montre la figure 6.16d, l'affinage ne permet pas au réseau d'estimer les variations de profondeurs sur les vidéos acquises avec une lampe à fente. Nos hypothèses sont que les images qui constituent la base ne sont pas assez diversifiées et présentent de trop faibles variations de profondeur pour envisager un apprentissage sur ce type de réseau.

En ce qui concerne les tests sur la base endoscopique, on observe à travers la comparaison des figures 6.16c et 6.15e qu'il y a bien eu une évolution du réseau. Les différences sont, faibles mais les contours semblent plus marqués, la zone utile ressort plus et l'outil est plus mis en avant. En revanche, le code couleur montre les incohérences de l'apprentissage. Celui-ci accorde toujours de l'importance à la zone à l'extérieur de la zone utile et lui trouve attribue toujours des variations de profondeur. De plus, les variations de profondeurs estimées dans la zone utile ne semblent pas cohérentes avec la structure concave de la rétine.

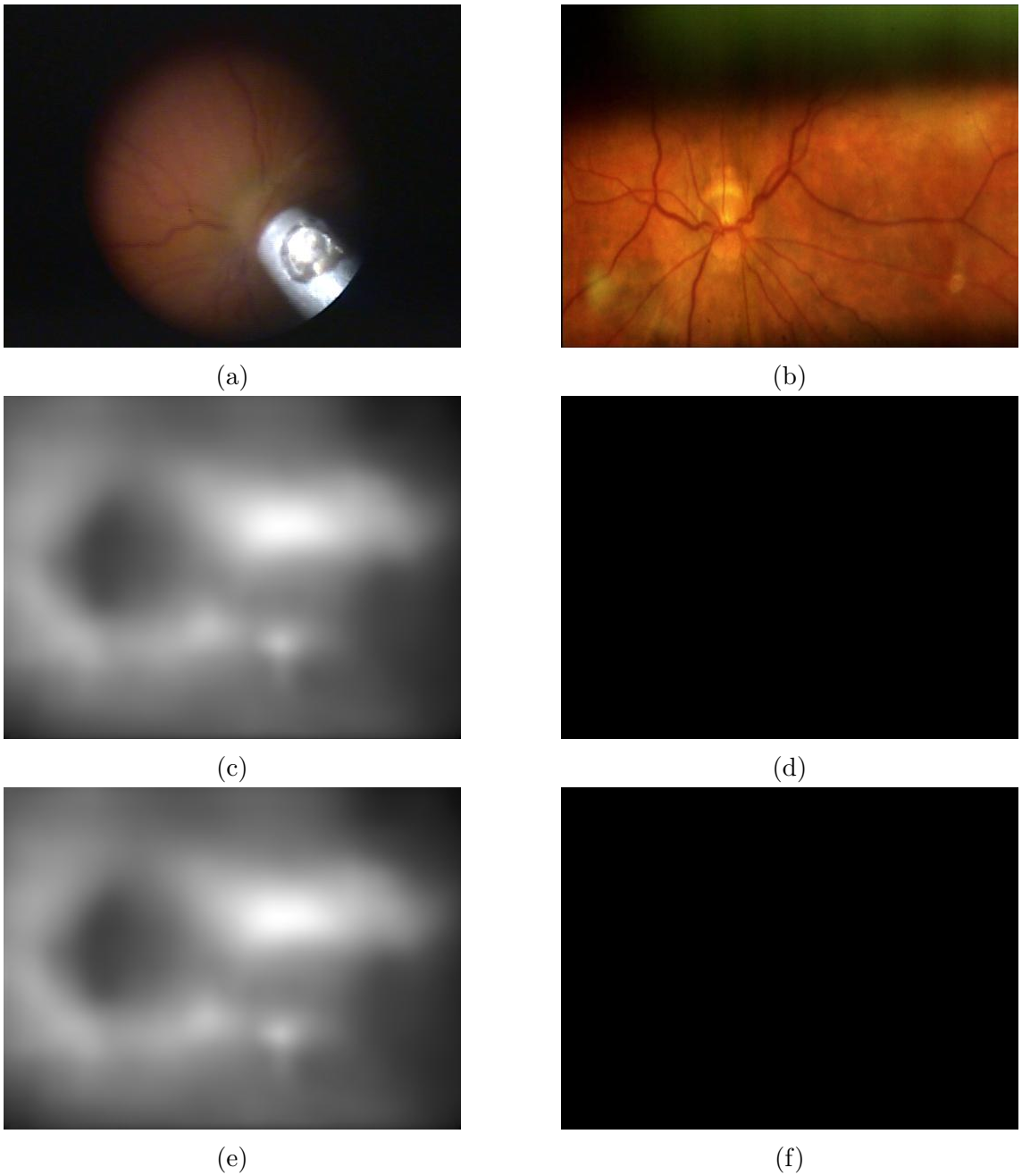
Dans un second temps, nous réalisons des affinages avec des versions normalisées des bases de données afin de tenter de palier aux grandes variations d'éclairages plutôt mal gérées par le réseau (selon [76] et [70]). Les paramétrages sont les mêmes que précédemment et les images 6.16e et 6.16f illustrent les résultats. Nous n'observons pas de changement pour le réseau affiné sur la base de vidéos acquises à la lampe à fente. En ce qui concerne l'endoscopie, les images 6.16e et 6.16c sont très proches. Les contours sont légèrement moins marqués et l'image est plus floue. Cet aspect se retrouve sur le reste de la base de test, aussi nous concluons que la normalisation n'améliore globalement pas les résultats et que d'autres solutions doivent être envisagées pour tirer profit de ce type de réseau estimant la profondeur à partir d'une image.

6.2.4 Bilan

Dans ce chapitre, nous avons vu que les CNN de types FlowNet peuvent donner des résultats encourageants quant à l'estimation sur nos bases rétiniennes. C'est d'ailleurs cette modalité qui donne les résultats les plus satisfaisants sur la base "lampe à fente" et la seule méthode à donner des estimations cohérentes sur la base "endoscope". Les résultats pour la création de mosaïques sont encourageants.

En ce qui concerne l'estimation de profondeur par le réseau SFM Learner, nous sommes arrivés à la conclusion que ce réseau n'est pas adapté à nos vidéos de rétines. Dans leur article [70], Rau et al. arrivent aux mêmes conclusions sur la modalité d'endoscopie digestive. Les principaux problèmes sont les variations d'éclairage et les textures qui ne sont pas assez variées au sein d'une même image.

Afin d'estimer la profondeur, ils ont recours à une méthode proche de celle que nous avons utilisé pour l'estimation des déplacements en faisant appel à un réseau à apprentissage fortement supervisé. Pour ce faire, ils ont dû créer artificiellement une base de données annotée de coloscopie en générant en 3D des colons. Ils se sont



(a) (b) (c) (d) (e) (f)

FIGURE 6.16: Comparaisons des cartes de profondeurs sur des images issues de nos bases de données à l'issue d'apprentissages par transfert. a. Image acquise à l'endoscope - b. Image acquise à la lampe à fente - c. Carte de profondeur estimée après apprentissage par transfert (KITTI (384×512px) puis base "endoscopie") - d. Carte de profondeur estimée après apprentissage par transfert (KITTI (384×512px) puis base "lampe à fente") - e. Carte de profondeur estimée après apprentissage par transfert (KITTI (384×512px) puis base "endoscopie" normalisée) - f. Carte de profondeur estimée après apprentissage par transfert (KITTI (384×512px) puis base "endoscopie" normalisée)

servis (entre-autre) de cette base simulée pour entraîner leur réseau et l'utiliser sur des données réelles.

7
Conclusion et discussion

Au cours de cette thèse, l'objectif était d'améliorer la qualité de vidéos de la chambre postérieure de l'œil afin de rendre les examens ou les chirurgies de cette zone plus confortables pour les médecins. Pour ce faire nous avons décidé de réaliser des cartes dynamiques et en 3 dimensions de rétines. Les deux outils d'acquisition d'images étudiés étaient la lampe à fente et l'endoscope oculaire. Afin de réaliser ces mosaïques d'images, les principales méthodes d'estimation de mouvement, entre deux images, de la littérature ont été testées. Nous les avons regroupées en deux catégories : les méthodes "classiques" et les méthodes utilisant l'apprentissage profond.

Les résultats des méthodes "classiques" testées sont variables et aucune n'a permis une estimation acceptable des déplacements pour la base de vidéos acquises à l'endoscope oculaire. En revanche, nous avons constaté que dans certains cas, les méthodes basées détection de points d'intérêt de type SURF pouvaient proposer des estimations convenables des déplacements entre deux images de vidéos acquises à la lampe à fente. Nous expliquons de tels résultats par la mauvaise qualité des vidéos acquises à l'endoscope. Celles-ci présentent des fortes variations d'éclairage d'une image à l'autre, mais également au sein d'une même image. En effet, il n'est pas rare de trouver des zones d'images extrêmement sombres à coté de zones complètement saturées par l'éclairage. De plus, les images sont très bruitées et présentent de faibles variations de textures. Les vidéos acquises à la lampe à fente sont également concernées par ces remarques, mais dans une moindre mesure.

Les méthodes de la seconde catégorie utilisent des réseaux de neurones convolutifs. On y retrouve notamment les réseaux de type FlowNet qui sont des réseaux à apprentissage fortement supervisé. Cette architecture nous a poussés à construire des bases de données de déplacements générés artificiellement et ayant pour fond la rétine afin d'optimiser les phases d'entraînement pour notre problématique. En effet, il n'existait pas de bases de données de vidéos rétiniennes ayant des annotations

pouvant servir de vérité terrain sur les déplacements entre chaque image.

Avec ces méthodes, les résultats sont plus concluants et plus précis qu'avec les méthodes classiques. Nous avons d'ailleurs pu réaliser des mosaïques en deux dimensions pour les deux types d'acquisition vidéo.

Enfin, un réseau à apprentissage auto-supervisé, ayant pour but d'estimer la profondeur d'une scène à partir d'une image et le déplacement de la caméra à partir de trois images, a été testé. Après plusieurs tests et des résultats en demi-teinte, il en est ressorti que ce type de réseau n'était pas adapté à nos données présentant des textures peu variées et de trop fortes variations d'éclairage. Une autre équipe travaillant sur l'endoscopie digestive est arrivée aux mêmes conclusions que nous et propose comme alternative une approche fortement supervisée utilisant notamment une base de données générées à partir d'un colon modélisé artificiellement en 3 dimensions.

Une amélioration possible de notre méthode pourrait donc être de générer artificiellement des rétines en 3 dimensions et de constituer notre base de données de vidéos rétiniennes afin d'entraîner un réseau similaire au leur. Ce faisant nous pourrions faire varier les textures, les éclairages et ajouter des outils à notre guise. Nous pourrions également quantifier aisément les résultats de nos entraînements.

Cette direction semble la plus prometteuse. Toutefois en restant sur le réseau SFM Learner nous pourrions tout de même forcer le masquage autour de la zone utile grâce à l'estimateur de zone utile que nous avons développé. Cependant, au vu des résultats du chapitre 6 et des conclusions de Rau et al., cette modification de ne devrait pas engendrer d'amélioration importante des résultats.

Nous pensons également qu'une amélioration de l'estimation des déplacements sur les vidéos acquises à la lampe à fente pourrait être faite en construisant une base d'entraînement plus spécifique aux caractéristiques de ce type d'acquisition. Comme pour la construction de Sliding Retinas I et II celle-ci pourrait puiser dans les fonds d'œil de la base du concours Kaggle. Enfin, une modification de Sliding Retinas II pourrait être faite pour être beaucoup plus spécifique à la modalité d'endoscopie oculaire en dégradant la qualité des images (ajout de saturations et de zones sombres) et en ajoutant artificiellement un outil mobile aux images.

En conclusion, nous avons proposé une solution pour agrandir le champs visuel dans les vidéos d'endoscopie oculaire et de lampe à fente. Au vu des résultats, la solution doit encore être améliorée. A terme, elle permettra un plus grand confort visuel du chirurgien et donc une intervention plus efficace et plus sûre.

8 Annexe

Cette partie se consacre à la délimitation de la zone utile dans les vidéos d'endoscopie oculaire. On appelle zone utile la zone du capteur sur laquelle se forme l'image. En effet, le capteur de l'endoscope est de forme rectangulaire tandis que le bouquet de fibres optique forme un signal plutôt circulaire/elliptique qui n'occupe pas la totalité du capteur.

8.1 Détection de la zone utile

Afin de ne concentrer les estimations des déplacements que sur la zone utile de l'image, nous avons décidé de mettre en place un masque pour la délimiter automatiquement dans le cadre des vidéos d'endoscopie oculaire. Comme nous l'avons vu, cette dernière peut se déplacer entre deux séquences et même au cours d'une séquence d'enregistrement. Il est donc important de pouvoir mettre à jour la position et la forme de ce masque.

Nous avons, dans un premier temps, délimité manuellement cette zone utile pour plusieurs images de la base de données d'enregistrement de chirurgies endoscopique. Elles sont visualisables sur la colonne de gauche de la figure 8.1. Ces zones sont assimilables à des ellipses, voilà pourquoi nous avons décidé de définir un modèle de masque elliptique. Le choix de l'ellipse s'est fait au profit de celui du cercle. En effet, en comparant les scores de valeur prédictive positive et de sensibilité d'ellipses et de cercles ajustés de manière guidée à la forme servant de vérité terrain nous obtenons 99.3 pourcents en valeur prédictive positive et 99 pourcents en sensibilité pour les ellipses et 88.1 pourcents en valeur prédictive positive et 98.2 pourcents en sensibilité pour les cercles. Le guidage s'est fait en donnant à l'algorithme d'estimation de forme, (développé dans la suite de cette section) l'information du contour vérité terrain du masque. La deuxième et la troisième colonne de la figure 8.1 montrent respectivement quelques exemples d'estimations guidées de cercles et d'ellipses.

En vert, on retrouve la zone commune au masque vérité terrain et au masque

estimé. Le rouge correspond à la zone du masque vérité terrain qui n'est pas détectée par la méthode et le bleu correspond à la zone du masque estimé qui n'appartient pas au masque vérité terrain. Sur la figure 8.1, la valeur prédictive positive se traduit par le quotient entre la zone verte et la zone verte plus la zone bleue. La sensibilité se traduit, quant à elle, par le quotient entre la zone verte et la zone verte plus la zone rouge.

Pour automatiser et ajuster au mieux le masque elliptique de manière automatique, nous développons une méthode qui calcule une image moyenne à partir de cent images consécutives de la vidéo et extrayons le canal vert pour faire ressortir au mieux la zone utile du fond. Le choix de la position et de la taille de l'ellipse optimale sont déterminés en maximisant la somme pondéré de deux fonction de coût. La première vise à maximiser le gradient du pourtour de l'ellipse et la seconde, la somme des intensité des pixels qui forment cette ellipse.

Sur les 2500 images annotées, la méthode automatique nous donne une valeur prédictive positive de 93.3 pourcents en moyenne et une sensibilité de 81.5 pourcents comme le montrent les troisièmes et quatrièmes colonnes de la figure 8.1 et le tableau 8.1.

Type de masque	Circulaire guidé	Elliptique guidé	Elliptique automatique
Val. prédict. pos. (%)	88.1	99.3	93.3
Sensibilité (%)	98.2	99.0	81.5

TABLE 8.1: Tableau récapitulatif des erreurs absolues moyennes pour chaque méthode

Ces résultats sont plutôt bons et satisfaisants pour une première phase d'expérimentation. En revanche, ils pourraient certainement être améliorés en entraînant un CNN à détecter automatiquement cette zone, par exemple.

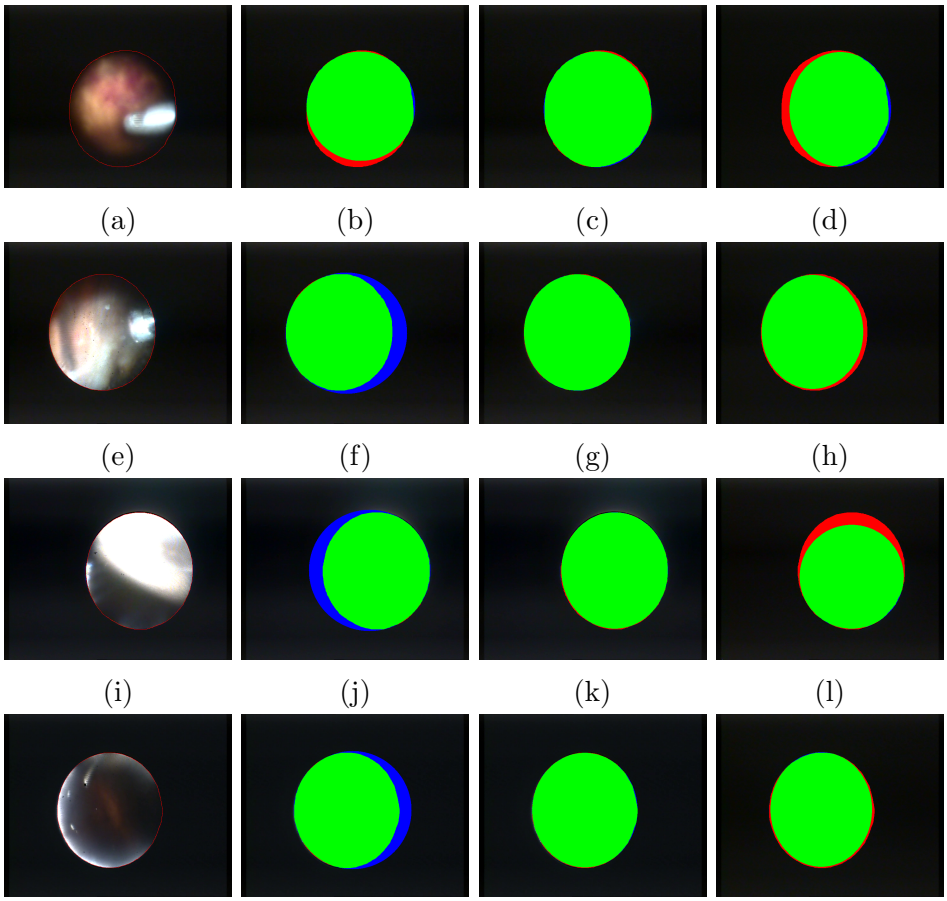


FIGURE 8.1: Exemples d'estimation de la zone utile en endoscopie oculaire. a. Exemple d'image avec contour masque vérité terrain - b. Meilleur ajustement guidé d'un masque circulaire - c. Meilleur ajustement guidé d'un masque elliptique - d. Meilleur ajustement automatique d'un masque elliptique - e. Exemple d'image avec contour masque vérité terrain - f. Meilleur ajustement guidé d'un masque circulaire - g. Meilleur ajustement guidé d'un masque elliptique - h. Meilleur ajustement automatique d'un masque elliptique - i. Exemple d'image avec contour masque vérité terrain - j. Meilleur ajustement guidé d'un masque circulaire - k. Meilleur ajustement guidé d'un masque elliptique - l. Meilleur ajustement automatique d'un masque elliptique - m. Exemple d'image avec contour masque vérité terrain - n. Meilleur ajustement guidé d'un masque circulaire - o. Meilleur ajustement guidé d'un masque elliptique - p. Meilleur ajustement automatique d'un masque elliptique

Bibliographie

[1] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox, "Flownet : Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[2] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "Flownet 2.0 : Evolution of optical flow estimation with deep networks," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, vol. 2, p. 6.

[3] Shmuel Peleg and Joshua Herman, "Panoramic mosaics by manifold projection," in *Proceedings of IEEE computer society conference on computer vision and pattern recognition*. IEEE, 1997, pp. 338–343.

[4] Shi Qiu, Dongmei Zhou, and Yun Du, "The image stitching algorithm based on aggregated star groups," *Signal, Image and Video Processing*, pp. 1–9, 2018.

[5] Li Li, Jian Yao, Yahui Liu, Wei Yuan, Shuzhu Shi, and Shenggu Yuan, "Optimal seamline detection for orthoimage mosaicking by combining deep convolutional neural network and graph cuts," *Remote Sensing*, vol. 9, no. 7, pp. 701, 2017.

[6] Abdul Qayyum, Aamir Saeed Malik, Naufal M Saad, Mahboob Iqbal, Mohd Faris Abdullah, Waqas Rasheed, Tuan AB Rashid Abdullah, and Mohd Yaqoob Bin Jafaar, "Scene classification for aerial images based on cnn using sparse coding technique," *International journal of remote sensing*, vol. 38, no. 8-10, pp. 2662–2685, 2017.

[7] Julian Colorado, Ivan Mondragon, Juan Rodriguez, and Carolina Castiblanco, "Geo-mapping and visual stitching to support landmine detection using a low-cost uav," *International Journal of Advanced Robotic Systems*, vol. 12, no. 9, pp. 125, 2015.

[8] Huajian Liu and Sang-Heon Lee, "Stitching of video sequences for weed mapping," in *2015 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*. IEEE, 2015, pp. 441–444.

[9] Jakub Ceranka, Mathias Polfliet, Frédéric Lecouvet, Nicolas Michoux, Johan de Mey, and Jef Vandemeulebroucke, "Registration strategies for multi-modal whole-body mri mosaicing," *Magnetic resonance in medicine*, vol. 79, no. 3, pp. 1684–1695, 2018.

[10] Lukasz Maciura and Jan G Bazan, "Granular computing in mosaicing of images from capsule endoscopy," *Natural computing*, vol. 14, no. 4, pp. 569–577, 2015.

[11] Daniel Reichard, Sebastian Bodenstedt, Stefan Suwelack, Benjamin Mayer, Anas Preukschas, Martin Wagner, Hannes Kenngott, Beat Müller-Stich, Rüdiger Dillmann, and Stefanie Speidel, "Intraoperative on-the-fly organ-mosaicking for laparoscopic surgery," *Journal of Medical Imaging*, vol. 2, no. 4, pp. 045001, 2015.

[12] Giovanni D De Palma, Stefania Staibano, Saverio Siciliano, Marcello Persico, Stefania Masone, Francesco Maione, Maria Siano, Massimo Mascolo, Dario Esposito, Francesca Salvatori, et al., "In vivo characterisation of superficial colorectal neoplastic lesions with high-resolution probe-based confocal laser endomicroscopy in combination with video-mosaicing : a feasibility study to enhance routine endoscopy," *Digestive and Liver Disease*, vol. 42, no. 11, pp. 791–797, 2010.

[13] Mireille Reeff, Friederike Gerhard, Philippe C Cattin, and Gábor Székely, "Mosaicing of endoscopic placenta images.," *GI Jahrestagung (1)*, vol. 2006, pp. 467–474, 2006.

[14] Pankaj Daga, François Chadebecq, Dzhoshkun I Shakir, Luis Carlos Garcia-Peraza Herrera, Marcel Tella, George Dwyer, Anna L David, Jan Deprest, Danail Stoyanov, Tom Vercauteren, et al., "Real-time mosaicing of fetoscopic videos using sift," in *Medical Imaging 2016 : Image-Guided Procedures, Robotic Interventions, and Modeling*. International Society for Optics and Photonics, 2016, vol. 9786, p. 97861R.

[15] Philippe C Cattin, Herbert Bay, Luc Van Gool, and Gábor Székely, "Retina mosaicing using local features," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2006, pp. 185–192.

[16] Jane Asmuth, Bojidar Madjarov, Paul Sajda, and Jeffrey W Berger, "Mosaicking and enhancement of slit lamp biomicroscopic fundus images," *British journal of ophthalmology*, vol. 85, no. 5, pp. 563–565, 2001.

[17] Rogério Richa, Rodrigo Linhares, Eros Comunello, Aldo Von Wangenheim, Jean-Yves Schnitzler, Benjamin Wassmer, Claire Guillemot, Gilles Thuret, Philippe Gain, Gregory Hager, et al., "Fundus image mosaicking for information augmentation in computer-assisted slit-lamp imaging," *IEEE transactions on medical imaging*, vol. 33, no. 6, pp. 1304–1312, 2014.

[18] Sandro De Zanet, Tobias Rudolph, Rogerio Richa, Christoph Tappeiner, and Raphael Sznitman, "Retinal slit lamp video mosaicking," *International journal of computer assisted radiology and surgery*, vol. 11, no. 6, pp. 1035–1041, 2016.

[19] James J Gibson, "The perception of the visual world," 1950.

[20] Bruce D Lucas, Takeo Kanade, et al., "An iterative image registration technique with an application to stereo vision," 1981.

[21] Berthold KP Horn and Brian G Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[22] Gunnar Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.

[23] A Danudibroto, Olivier Gérard, Martino Alessandrini, Oana Mirea, Jan D'hooge, and Eigil Samset, "3d farnebäck optic flow for extended field of view of echocardiography," in *International Conference on Functional Imaging and Modeling of the Heart*. Springer, 2015, pp. 129–136.

[24] Naoki Chiba, Hiroshi Kano, Michihiko Minoh, and Masashi Yasuda, "Feature-based image mosaicing," *Systems and Computers in Japan*, vol. 31, no. 7, pp. 1–9, 2000.

[25] Stanley T Birchfield and Shrinivas J Pundlik, "Joint tracking of features and edges," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–6.

[26] Andrés Bruhn, Joachim Weickert, and Christoph Schnörr, "Lucas/kanade meets horn/schunck : Combining local and global optic flow methods," *International journal of computer vision*, vol. 61, no. 3, pp. 211–231, 2005.

[27] Dinh Hoan Trinh, Christian Daul, Walter Blondel, and Dominique Lamarque, "Mosaicing of images with few textures and strong illumination changes : Application to gastroscopic scenes," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1263–1267.

[28] Yinlin Hu, Rui Song, and Yunsong Li, "Efficient coarse-to-fine patchmatch for large displacement optical flow," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5704–5712.

[29] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman, "Patchmatch : A randomized correspondence algorithm for structural image editing," in *ACM Transactions on Graphics (ToG)*. ACM, 2009, vol. 28, p. 24.

[30] Jaswant Jain and Anil Jain, "Displacement measurement and its application in interframe image coding," *IEEE Transactions on communications*, vol. 29, no. 12, pp. 1799–1808, 1981.

[31] J0 Limb and J Murphy, "Measuring the speed of moving objects from television signals," *IEEE Transactions on Communications*, vol. 23, no. 4, pp. 474–478, 1975.

[32] Fabio Rocca and Silvio Zanoletti, "Bandwidth reduction via movement compensation on a model of the random video process," *IEEE Transactions on Communications*, vol. 20, no. 5, pp. 960–965, 1972.

[33] Thomas Komarek and Peter Pirsch, "Array architectures for block matching algorithms," *IEEE Transactions on Circuits and Systems*, vol. 36, no. 10, pp. 1301–1308, 1989.

[34] Aroh Barjatya, "Block matching algorithms for motion estimation," *IEEE Transactions Evolution Computation*, vol. 8, no. 3, pp. 225–239, 2004.

[35] Reoxiang Li, Bing Zeng, and Ming L Liou, "A new three-step search algorithm for block motion estimation," *IEEE transactions on circuits and systems for video technology*, vol. 4, no. 4, pp. 438–442, 1994.

[36] Lai-Man Po and Wing-Chung Ma, "A novel four-step search algorithm for fast block motion estimation," *IEEE transactions on circuits and systems for video technology*, vol. 6, no. 3, pp. 313–317, 1996.

[37] Yao Nie and Kai-Kuang Ma, "Adaptive rood pattern search for fast block-matching motion estimation," *IEEE Transactions on Image processing*, vol. 11, no. 12, pp. 1442–1449, 2002.

[38] Nehal N Shah and Upena D Dalal, "Hardware efficient double diamond search block matching algorithm for fast video motion estimation," *Journal of Signal Processing Systems*, vol. 82, no. 1, pp. 115–135, 2016.

[39] Shan Zhu and Kai-Kuang Ma, "A new diamond search algorithm for fast block-matching motion estimation," *IEEE transactions on Image Processing*, vol. 9, no. 2, pp. 287–290, 2000.

[40] Tom Vercauteren, Aymeric Perchant, Xavier Pennec, and Nicholas Ayache, "Mosaicing of confocal microscopic in vivo soft tissue video sequences," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2005, pp. 753–760.

[41] Frédéric Dufaux and Fabrice Moscheni, "Background mosaicking for low bit rate video coding," in *Proceedings of 3rd IEEE International Conference on Image Processing*. IEEE, 1996, vol. 1, pp. 673–676.

[42] Minh Tân Pham and Didier Gueriot, "Guided block-matching for sonar image registration using unsupervised kohonen neural networks," in *2013 OCEANS-San Diego*. IEEE, 2013, pp. 1–5.

[43] Cyril Chailloux, Jean-Marc Le Caillec, Didier Gueriot, and Benoit Zerr, "Intensity-based block matching algorithm for mosaicing sonar images," *IEEE Journal of Oceanic Engineering*, vol. 36, no. 4, pp. 627–645, 2011.

[44] Hans P Moravec, "Rover visual obstacle avoidance.," in *IJCAI*, 1981, pp. 785–790.

[45] David G Lowe, "Object recognition from local scale-invariant features," in *iccv*. Ieee, 1999, p. 1150.

[46] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf : Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[47] Deepak Geetha Viswanathan, "Features from accelerated segment test (fast)," 2009.

[48] Martin A Fischler and Robert C Bolles, "Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[49] Filip Schouwenaars, Radu Timofte, and Luc J Van Gool, "Robust scene stitching in large scale mobile mapping.," in *BMVC*, 2013.

[50] Wen Rong, Hui Chen, Jiaju Liu, Yanyan Xu, and Ralf Haeusler, "Mosaicing of microscope images based on surf," in *2009 24th International Conference Image and Vision Computing New Zealand*. IEEE, 2009, pp. 271–275.

[51] Valentin Becker, Tom Vercauteren, Claus Hann von Weyhern, Christian Prinz, Roland M Schmid, and Alexander Meining, "High-resolution miniprobe-based confocal microscopy in combination with video mosaicing (with video)," *Gastrointestinal endoscopy*, vol. 66, no. 5, pp. 1001–1007, 2007.

[52] Kunihiko Fukushima, "Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[53] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, "Learning internal representations by error propagation," Tech. Rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[54] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[55] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[56] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[57] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

[58] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid, "Epicflow : Edge-preserving interpolation of correspondences for optical flow," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1164–1172.

[59] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid, "Deepflow : Large displacement optical flow with deep matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1385–1392.

[60] Yann LeCun, "The mnist database of handwritten digits," *http ://yann. lecun. com/exdb/mnist/*, 1998.

[61] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. IEEE, 2006, vol. 2, pp. 1735–1742.

[62] Qi Zhao, Boxue Zhang, Shuchang Lyu, Hong Zhang, Daniel Sun, Guoqiang Li, and Wenquan Feng, "A cnn-sift hybrid pedestrian navigation method based on first-person vision," *Remote Sensing*, vol. 10, no. 8, pp. 1229, 2018.

[63] Floris Gaisser, Pieter P Jonker, and Toshio Chiba, "Image registration for placenta reconstruction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 33–40.

[64] D Butler, Jonas Wulff, G Stanley, and M Black, "Mpi-sintel optical flow benchmark : Supplemental material," in *MPI-IS-TR-006, MPI for Intelligent Systems (2012*. Citeseer, 2012.

[65] Matthew D Zeiler, Graham W Taylor, Rob Fergus, et al., "Adaptive deconvolutional networks for mid and high level feature learning.," in *ICCV*, 2011, vol. 1, p. 6.

[66] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.

[67] Laura Sevilla-Lara, Deqing Sun, Varun Jampani, and Michael J Black, "Optical flow with semantic segmentation and localized layers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3889–3898.

[68] Dinghuang Ji, Junghyun Kwon, Max McFarland, and Silvio Savarese, "Deep view morphing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2155–2163.

[69] Paul Wohlhart and Vincent Lepetit, "Learning descriptors for object recognition and 3d pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3109–3118.

[70] Anita Rau, PJ Eddie Edwards, Omer F Ahmad, Paul Riordan, Mirek Janatka, Laurence B Lovat, and Danail Stoyanov, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *International journal of computer assisted radiology and surgery*, vol. 14, no. 7, pp. 1167–1176, 2019.

[71] Kishore Konda and Roland Memisevic, "Unsupervised learning of depth and motion," *arXiv preprint arXiv :1312.3429*, 2013.

[72] Viorica Patraucean, Ankur Handa, and Roberto Cipolla, "Spatio-temporal video autoencoder with differentiable memory," *arXiv preprint arXiv :1511.06309*, 2015.

[73] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis, "Back to basics : Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–10.

[74] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid, "Unsupervised cnn for single view depth estimation : Geometry to the rescue," in *European Conference on Computer Vision*. Springer, 2016, pp. 740–756.

[75] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu, "Occlusion aware unsupervised learning of optical flow," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4884–4893.

[76] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1851–1858.

[77] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox, "Demon : Depth and motion network for learning monocular stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5038–5047.

[78] Zhichao Yin and Jianping Shi, "Geonet : Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1983–1992.

[79] Simon Meister, Junhwa Hur, and Stefan Roth, "Unflow : Unsupervised learning of optical flow with a bidirectional census loss," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[80] Mohammad Ali Armin, Nick Barnes, Salman Khan, Miaomiao Liu, Florian Grimpen, and Olivier Salvado, "Unsupervised learning of endoscopy video frames' correspondences from global and local transformation," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pp. 108–117. Springer, 2018.

[81] Frank Rosenblatt, "The perceptron : a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, pp. 386, 1958.

[82] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato, "Phrase-based & neural unsupervised machine translation," *arXiv preprint arXiv :1804.07755*, 2018.

[83] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[84] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning : A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[85] Allvar Gullstrand, "Demonstration der nernstspaltlampe," *Ber Deutsch Ophthalmol Ges*, vol. 37, pp. 374–376, 1911.

[86] Marcus-Matthias Gellrich, "The fundus slit lamp," *SpringerPlus*, vol. 4, no. 1, pp. 56, 2015.

[87] Adolf F Fercher, Hai C Li, and Christoph K Hitzenberger, "Slit lamp laser doppler interferometer," *Lasers in Surgery and Medicine*, vol. 13, no. 4, pp. 447–452, 1993.

[88] Hans Hoerauf, Christopher Wirbelauer, Christian Scholz, Ralf Engelhardt, Peter Koch, Horst Laqua, and Reginald Birngruber, "Slit-lamp-adapted optical coherence tomography of the anterior segment," *Graefe's archive for clinical and experimental ophthalmology*, vol. 238, no. 1, pp. 8–18, 2000.

[89] Christopher Wirbelauer, Christian Scholz, Hans Hoerauf, Duy Thoai Pham, Horst Laqua, and Reginald Birngruber, "Noncontact corneal pachymetry with

slit lamp-adapted optical coherence tomography," *American journal of ophthalmology*, vol. 133, no. 4, pp. 444–450, 2002.

[90] Robert A Moses, "The goldmann applanation tonometer," *American journal of ophthalmology*, vol. 46, no. 6, pp. 865–869, 1958.

[91] Gilbert Baum and Ivan Greenwood, "The application of ultrasonic locating techniques to ophthalmology : Ii. ultrasonic slit lamp in the ultrasonic visualization of soft tissues," *AMA Archives of ophthalmology*, vol. 60, no. 2, pp. 263–279, 1958.

[92] Marcus-Matthias Gellrich, *The slit lamp : applications for biomicroscopy and videography*, Springer Science & Business Media, 2013.

[93] Kristina Prokopetc and Adrien Bartoli, "A comparative study of transformation models for the sequential mosaicing of long retinal sequences of slit-lamp images obtained in a closed-loop motion," *International journal of computer assisted radiology and surgery*, vol. 11, no. 12, pp. 2163–2172, 2016.

[94] I Coghill, KC Jordan, RA Black, IAT Livingstone, and ME Giardini, "3d reconstruction of the fundus of a phantom eye through stereo imaging of slit lamp images," in *BioMedEng18*, 2018, pp. 216–216.

[95] Amol B Jagadale and DV Jadhav, "Early detection and categorization of cataract using slit-lamp images by hough circular transform," in *2016 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2016, pp. 0232–0235.

[96] Xiyang Liu, Jiewei Jiang, Kai Zhang, Erping Long, Jiangtao Cui, Mingmin Zhu, Yingying An, Jia Zhang, Zhenzhen Liu, Zhuoling Lin, et al., "Localization and diagnosis framework for pediatric cataracts based on slit-lamp images using deep features of a convolutional neural network," *PloS one*, vol. 12, no. 3, pp. e0168606, 2017.

[97] Andrea Russo, Francesco Morescalchi, Ciro Costagliola, Luisa Delcassi, and Francesco Semeraro, "Comparison of smartphone ophthalmoscopy with slit-lamp biomicroscopy for grading diabetic retinopathy," *American journal of ophthalmology*, vol. 159, no. 2, pp. 360–364, 2015.

[98] Harvey E Thorpe et al., "Ocular endoscope : instrument for removal of intravitreous non magnetic foreign bodies," *Trans Am Acad Ophthalmol Otolaryngol*, vol. 39, pp. 422–424, 1934.

[99] Robert F Butterworth and John L Bignell, "A new type of eye endoscope," *The British journal of ophthalmology*, vol. 36, no. 4, pp. 217, 1952.

[100] John L Norris and Gilbert W Cleasby, "An endoscope for ophthalmology," *American journal of ophthalmology*, vol. 85, no. 3, pp. 420–422, 1978.

[101] Veniamin V Volkov, Andrey V Danilov, Leonid N Vassin, and Yurii A Frolov, "Flexible endoscopes : Ophthalmoendoscopic techniques and case reports," *Archives of Ophthalmology*, vol. 108, no. 7, pp. 956–957, 1990.

[102] Shuichiro Eguchi and Makoto Araie, "A new ophthalmic electronic videoendoscope system for intraocular surgery," *Archives of ophthalmology*, vol. 108, no. 12, pp. 1778–1781, 1990.

[103] Martin Uram, "Ophthalmic laser microendoscope endophotocoagulation," *Ophthalmology*, vol. 99, no. 12, pp. 1829–1832, 1992.

[104] Kyle V Marra, Yoshihiro Yonekawa, Thanos D Papakostas, and Jorge G Arroyo, "Indications and techniques of endoscope assisted vitrectomy," *Journal of ophthalmic & vision research*, vol. 8, no. 3, pp. 282, 2013.

[105] Shinichi Kawashima, Motoko Kawashima, and Kazuo Tsubota, "Endoscopy-guided vitreoretinal surgery," *Expert review of medical devices*, vol. 11, no. 2, pp. 163–168, 2014.

[106] Martin Uram, "Endoscopic cyclophotocoagulation in glaucoma management.," *Current opinion in ophthalmology*, vol. 6, no. 2, pp. 19–29, 1995.

[107] C Valmaggia and M De Smet, "Endoscopic laser coagulation of the ciliary processes in patients with severe chronic glaucoma," *Klinische Monatsblätter für Augenheilkunde*, vol. 221, no. 05, pp. 343–346, 2004.

[108] John E Moore, Gehan D Herath, and Anant Sharma, "Continuous curvilinear capsulorhexis with use of an endoscope," *Journal of Cataract & Refractive Surgery*, vol. 30, no. 5, pp. 960–963, 2004.

[109] Khalid Al Sabti, Seemant Raizada, and Talal Al AbdulJalil, "Cataract surgery assisted by anterior endoscopy," *British Journal of Ophthalmology*, vol. 93, no. 4, pp. 531–534, 2009.

[110] Claude Boscher and Ferenc Kuhn, "An endoscopic overview of the anterior vitreous base in retinal detachment and anterior proliferative vitreoretinopathy," *Acta ophthalmologica*, vol. 92, no. 4, pp. e298–e304, 2014.

[111] Shin Yoshitake, Hideyasu Oh, and Mihori Kita, "Endoscope-assisted vitrectomy for retinal detachment in an eye with microcornea," *Japanese journal of ophthalmology*, vol. 56, no. 6, pp. 613–616, 2012.

[112] Khalid Al Sabti and Seemant Raizada, "Endoscope-assisted pars plana vitrectomy in severe ocular trauma," *British Journal of Ophthalmology*, vol. 96, no. 11, pp. 1399–1403, 2012.

[113] Dal W Chun, Marcus H Colyer, and Keith J Wroblewski, "Visual and anatomic outcomes of vitrectomy with temporary keratoprosthesis or endoscopy in ocular trauma with opaque cornea," *Ophthalmic Surgery, Lasers and Imaging Retina*, vol. 43, no. 4, pp. 302–310, 2012.

[114] Joshua Ben-nun, "Cornea sparing by endoscopically guided vitreoretinal surgery," *Ophthalmology*, vol. 108, no. 8, pp. 1465–1470, 2001.

[115] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving ? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[116] Mathieu Aubry, Daniel Maturana, Alexei A Efros, Bryan C Russell, and Josef Sivic, "Seeing 3d chairs : exemplar part-based 2d-3d alignment using a large dataset of cad models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3762–3769.

[117] Hans Knutsson and C-F Westin, "Normalized and differential convolution," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1993, pp. 515–523.

[118] KJA Westin, AV Boiko, BGB Klingmann, VV Kozlov, and PH Alfredsson, "Experiments in a boundary layer subjected to free stream turbulence. part 1. boundary layer structure and receptivity," *Journal of Fluid Mechanics*, vol. 281, pp. 193–218, 1994.

[119] Gunnar Farnebäck, *Polynomial expansion for orientation and motion estimation*, Ph.D. thesis, Linköping University Electronic Press, 2002.

[120] Donald W Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[121] Olivier D Faugeras, Q-T Luong, and Stephen J Maybank, "Camera self-calibration : Theory and experiments," in *European conference on computer vision*. Springer, 1992, pp. 321–334.

[122] Rud Sturm, "Das problem der projectivität und seine anwendung auf die flächen zweiten grades," *Mathematische Annalen*, vol. 1, no. 4, pp. 533–574, 1869.

[123] Guido Hauck, "Neue constructionen der perspective und photogrammetrie.(theorie der trilinearen verwandtschaft ebener systeme, i. artikel.)," *Journal für die reine und angewandte Mathematik*, vol. 95, pp. 1–35, 1883.

[124] Stephen J Maybank and Olivier D Faugeras, "A theory of self-calibration of a moving camera," *International journal of computer vision*, vol. 8, no. 2, pp. 123–151, 1992.

[125] Q-T Luong and Olivier D Faugeras, "Self-calibration of a moving camera from point correspondences and fundamental matrices," *International Journal of computer vision*, vol. 22, no. 3, pp. 261–289, 1997.

[126] Elsayed E Hemayed, "A survey of camera self-calibration," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.* IEEE, 2003, pp. 351–357.

UNIVERSITE BIOLOGIE BRETAGNE SANTE LOIRE

UBO Université de Bretagne Occidentale

Titre : Champ visuel augmenté pour exploration vidéo de la rétine

Mots clés : Vidéos, rétine, mosaïque, estimation de déplacements, flux optique, réseau de neurones convolutifs

Résumé :
L'objectif de cette thèse est d'augmenter le confort visuel de l'ophtalmologue au cours d'examens ou de chirurgies de la rétine. Pour ce faire, nous décidons d'augmenter artificiellement et en temps réel le champ visuel dans le cas de vidéos d'exploration acquises à la lampe à fente et à l'endoscope oculaire. L'augmentation passe par la mise en place de cartes dynamiques en 3D de la rétine. A notre connaissance, il n'existe pas de telle méthode dans littérature.
Notre solution passe par l'étude de différentes méthodes d'estimation de déplacements entre deux images. Nous les regroupons en méthodes « classiques » d'une part, comptant notamment des méthodes basées sur les algorithmes SIFT ou SURF. D'autre part, nous rassemblons des méthodes utilisant l'apprentissage profond (ou méthodes « CNN » pour Convolutional Neural Network).
Certaines de ces méthodes, comme celles utilsant les réseaux FlowNet, nécessitent une annotation vérité terrain des déplacements entre image.
Comme de telles bases de données n'existent pas en ophtalmologie, des bases généralistes ont été utilisées. De plus, nous avons construit deux bases de données de déplacements artificiels ayant pour fond des images de rétines. Enfin, pour contourner le problème d'annotation, une approche utilisant l'apprentissage auto-supervisé a été étudiée.
Après comparaisons des résultats, il apparaît que les méthodes « CNN » surpassent les méthodes classiques. De plus, seule une supervision forte de l'apprentissage permet des résultats satisfaisants. A l'avenir, nous espérons que ces travaux pourront permettre aux chirurgiens d'être plus confiants et efficaces dans des environnements où il peut être compliqué de se repérer.

Title : Augmented field of view for videos of retinal exploration

Keywords : Videos, retina, mosaic, motion estimation, optical flow, convolutional neural network

Abstract :
The main objective of this thesis is to increase the visual comfort of the ophthalmologists during examinations or surgeries. To do so, we decided to artificially increase in real time the field of view in videos of retinal exploration. The tools used for the acquisition of these videos are the slit lamp and the endoscope. The increase of the field of view passes by the establishment of dynamic 3D maps of the retina.
To our knowledge, there is still no such method in the state of the art.
In order to implement our solution, we studied the different methods of motion estimations between two images. We grouped them into "classical" methods, on the one hand, including methods based on SIFT or SURF algorithms. On the other hand, we grouped deep learning methods (or "CNN" methods for Convolutional Neural Network).
Some of these methods, such as those using FlowNet networks, required ground truth annotation of movement between images.
Since such bases are very difficult to set up in the medical field and do not exist in ophthalmology, general databases have been used. In addition, we built two databases of artificial displacements which backgrounds are composed of images of retinas. Finally, to get around this problem of annotations, a self-supervised deep learning approach was studied.
After comparing the results, it appears that methods using convolutional neural networks outperform conventional methods for estimating movements in retinal videos. Moreover, only a strong supervision allows acceptable results. In the future, we hope that this work will enable surgeons to be more confident and effective in environments where it is sometimes difficult to find their bearings.