ARCHITECTURE ET CONCEPTION DE RETINES CMOS :INTEGRATION DE LA MESURE DU MOUVEMENT GLOBALDANS UN IMAGEUR

Fabrice Gensolen

UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

T H E S E

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

Discipline : Electronique, Optronique et Systèmes
Formation Doctorale : Systèmes Automatiques et Microélectroniques
Ecole Doctorale : Informations Structures Systèmes

présentée et soutenue publiquement

par

Fabrice GENSOLEN

le 25 septembre 2006

ARCHITECTURE ET CONCEPTION DE RETINES CMOS : INTEGRATION DE LA MESURE DU MOUVEMENT GLOBAL DANS UN IMAGEUR

JURY

M. François BERRY	Maître de Conférences à l'Université de Clermont-Ferrand	Examinateur
M. Guy CATHEBRAS	Maître de Conférences à l'Université de Montpellier II	Encadrant de Thèse
M. Pascal FOUILLAT	Professeur à l'Université de Bordeaux	Rapporteur
M. Lionel MARTIN	Ingénieur chez STMicroelectronics	Encadrant de Thèse
M. Michel ROBERT	Professeur à l'Université de Montpellier II	Directeur de Thèse
M. Frédéric TRUCHETET	Professeur à l'Université de Bourgogne	Rapporteur
M. Matteo VALENZA	Professeur à l'Université de Montpellier II	Président du Jury

A toute ma famille, dont l'amour et les efforts m'ont porté jusqu'ici…

REMERCIEMENTS

Les travaux présentés dans ce mémoire ont été menés dans le cadre d'une collaboration entre le LIRMM et la société STMicroelectronics (site de Rousset).

Aussi, je souhaiterais remercier tout d'abord Michel Robert (directeur du LIRMM), Michel Habib (ancien directeur), ainsi que Klaus Rischmuller et Bernard Kasser (responsables de l'équipe Advanced System Technology de STMicroelectronics Rousset) pour m'avoir accueilli dans leurs murs et y effectuer ma thèse.

En plus de ses fonctions à la tête du laboratoire, Michel Robert a également encadré ma thèse. Je lui adresse un remerciement particulier pour cela, ainsi que pour m'avoir fait bénéficier de sa vision scientifique, de son enthousiasme et de ses qualités de meneur d'hommes. Celles-ci m'ont été souvent bien utiles, notamment pour aller… droit au but !

Ensuite mes remerciements vont aux membres du jury : Pascal Fouillat et Frédéric Truchetet qui m'ont fait l'honneur de juger ma thèse ; François Berry, examinateur, qui m'a conseillé lors de nos différentes rencontres, ainsi que Matteo Valenza qui a présidé le jury.

Je tiens à remercier tout particulièrement Guy Cathebras, co-encadrant de ma thèse, qui m'a fait partager sa passion pour les sciences et la microélectronique durant ces années et dont les qualités pédagogiques et la rigueur scientifique en font un exemple pour moi. Je tiens également à saluer Isabelle, son épouse, pour sa patience lors des soirées de travail tardives…

Je remercie aussi bien vivement Lionel Martin, co-encadrant de ma thèse, pour son esprit critique lors de nos points d'avancement et pour ses conseils de rigueur et de méthode de travail. Ces derniers étaient d'autant plus utiles pour collaborer à distance…

J'exprime ma reconnaissance à toutes les personnes du département microélectronique du LIRMM, à toute l'équipe AST de ST Rousset et la division Imaging Grenoble (Pascal et Yvon notamment). Je les remercie pour leur disponibilité et leurs conseils (techniques et autres..), pour les moments partagés au bureau (hein Arnaud !) et ailleurs… lors des missions, « rigolades », barbecs, repas, sortie voilier, activités subaquatiques….. vraiment top !

Un grand merci également aux collègues roboticiens du LIRMM pour leur patience lors de mes questions de novice en traitement d'images… ainsi qu'à l'ensemble du personnel technique et administratif.

J'adresse aussi mes remerciements aux étudiants que j'ai pu encadrer durant leur stages ou projets (Stéphane, Alex, Greg et Fred, Bertrand, Cathy, Pierre-Olivier), et qui ont donc contribué à ce travail.

J'ai une pensée particulière pour tous les doctorants que j'ai eu l'occasion de croiser pendant ces années de thèse ! et j'en ai vu….. ☺☺☺ Outre les foots, volleys, et autres apéros, je me souviendrai longtemps de ces relais 4x2km lors des fêtes du sport annuelles à la fac, notamment celui que l'on a remporté !!!... à bientôt !!

Enfin, je ne peux terminer sans un clin d'oeil à ma Nine qui m'a vraiment bien aidé lors de la relecture (vive les 3 heures du mat !!)… et qui m'a permis de mieux gérer les grands moments… de « pression » !!!

Merci et excellente continuation à tous !!!

SOMMAIRE

INTRODUCTION GENERALE

Au début des années quatre-vingt-dix, les capteurs d'images CMOS n'étaient envisagés que dans le cadre de recherches scientifiques, la technologie CCD dominait alors. L'évolution extraordinaire des procédés de fabrication CMOS a fait qu'aujourd'hui ils atteignent la moitié des parts du marché. Ceci est également lié à l'avènement des dispositifs portables grand public tels que les téléphones mobiles, qui sont produits en très grand volume et pour lesquels le procédé CMOS est le mieux placé pour des raisons de coût d'intégration. En effet, la plupart des téléphones mobiles embarquent désormais les fonctions photo et/ou vidéo[1], et ils constituent un marché particulièrement attractif puisqu'il a représenté en 2005 un nombre de 700 millions d'unités vendues, tout en poursuivant sa croissance entamée il y a une dizaine d'année [IC-Insights05].

La conception de ces appareils portables doit respecter plusieurs contraintes fortes tel que le faible coût (~10 $), la faible consommation pour une autonomie maximum (~100mW) et la haute qualité. La réalisation de ces micro-capteurs d'imagerie intégrés constitue déjà une prouesse technologique en intégrant des technologies hétérogènes dans un même module.

D'un point de vue technologique, les progrès réalisés dans le domaine de l'acquisition des images sont considérables puisque l'on atteint des tailles de pixels de 2,25 µm de côté, alors que Eric Fossum [Fossum-97] indiquait en 1997 que la limite optique imposait au mieux un pas de 5 µm environ. La taille des pixels proche de 2 µm de côté constitue une limite en l'état actuel de la technologie optique [Rhodes et al.-04]. Ainsi, de nombreuses recherches portent aujourd'hui sur ces aspects en proposant de nouveaux matériaux [Video/ImagingDesignLine-06]. Cette fonction optique s'intègre dans des dispositifs électro-optiques toujours plus complexes (cf. page Annexe 1) dont les bonnes performances résultent d'une approche de conception « système ». En effet, la fonction d'acquisition (l'électronique et l'optique) et le post-traitement (amélioration du rendu et de la qualité d'image) sont interdépendantes.

On peut classer les circuits photosensibles en deux catégories : les imageurs et les rétines.

- Les imageurs sont exclusivement dédiés à l'acquisition de l'image, et à sa numérisation (cf. Figure 2). Dans certains cas, il est possible de paramétrer la stratégie d'adressage ou de lecture des pixels. Les paramètres peuvent alors être: le temps d'intégration des pixels, la vitesse de lecture, ou la zone de l'image acquise. Cette flexibilité s'avère par exemple intéressante dans le cadre de la poursuite d'objets dans une scène où la fenêtre d'intérêt est variable.
- Les rétines ou circuits de vision effectuent un traitement sur l'information lumineuse acquise pour en extraire une information, qui n'est pas nécessairement l'image. Ils sont alors :
  - soit dédiés, pour répondre à une application ou un besoin spécifique. Un exemple de ce type de systèmes est la souris optique [Tanner & Mead-86].

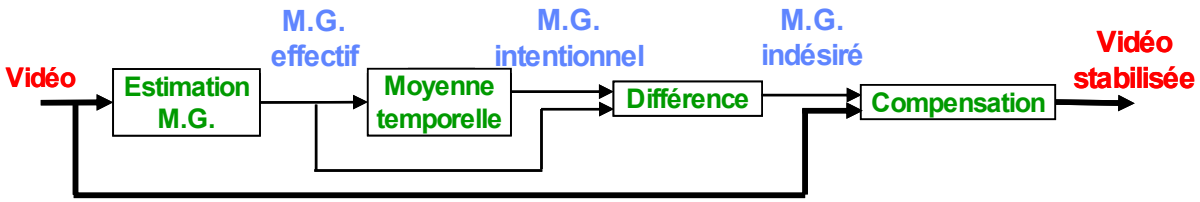
[1] Pour se démarquer sur ce marché, certains fabricants (Nokia, Samsung, Sanyo) proposent même de nouveaux appareils destinés à la 3eme génération, équipés de deux imageurs : l'un dédié à la photo et l'autre à la visiophonie en se situant sur la face avant du mobile.

- soit généralistes, et programmables pour pouvoir ainsi réaliser des tâches diverses souvent basées sur des architectures SIMD (Single Instruction Multiple Data). Un exemple est la rétine programmable de deuxième génération [Paillet et al.-98].

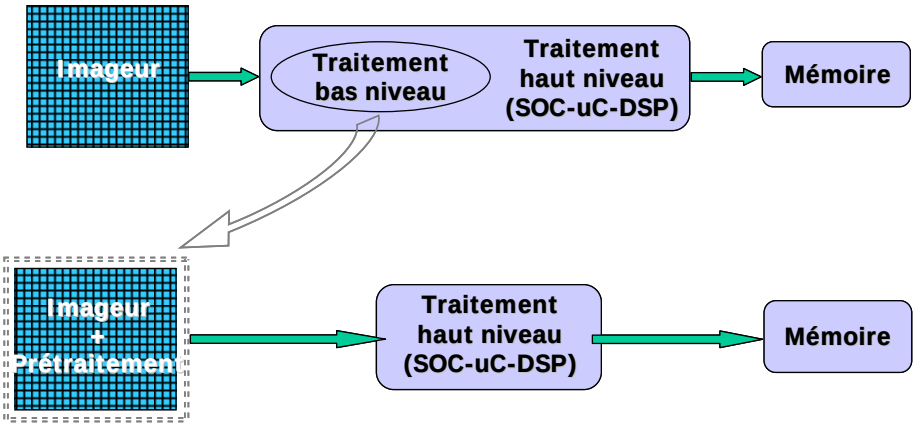
Notre travail de recherche se positionne dans ce contexte des capteurs d'images et des systèmes de vision, avec pour objectif d'intégrer de nouvelles fonctionnalités à forte valeur ajoutée aux imageurs fabriqués par la société STMicroelectronics. En effet, une entreprise qui conçoit et fabrique de tels capteurs, destinés à être embarqués dans des dispositifs mobiles tels que les téléphones portables ou les assistants numériques personnels (« PDA »), souhaite ainsi se démarquer de la concurrence et augmenter ses parts de marché.

Dans le cadre de ces travaux, nous nous intéressons à l'ajout de la stabilisation vidéo. En effet, cette fonction s'avère nécessaire lors de la prise de vue à l'aide de dispositifs portables, très sujets aux tremblements.

La stabilisation électronique consiste à ''filtrer'' ces mouvements non voulus de la caméra pour ne conserver et ne restituer que le mouvement intentionnel, de type panoramique par exemple. L'opération se décompose alors en quatre étapes, décrites sur la figure ci-dessous :



La première étape de l'estimation du mouvement global inter images représente entre 50% et 70% de la charge de calcul globale pour réaliser la stabilisation vidéo. Ainsi, le challenge auquel nous nous attachons est de reporter cette tâche, ou une partie importante de celle-ci, au niveau du plan focal afin de réduire la charge calculatoire associée. Finalement, notre capteur devra embarquer une double fonctionnalité : l'acquisition vidéo classique à 30 images/s et l'estimation du mouvement global entre deux images consécutivement acquises.



Malheureusement ces deux fonctions sont contradictoires en termes d'intégration au niveau pixel et il y a un compromis à trouver. En effet, il faut associer l'acquisition d'image, dont le critère majeur de performance est la résolution de l'image (donc des pixels petits), avec le traitement dans le plan focal de l'information lumineuse acquise pour en extraire le mouvement qui implique l'ajout de transistors au sein du pixel (donc des pixels plus gros).

Les technologies TFA ou Above IC, qui dissocient la fonction "transduction" du substrat en déposant une couche photosensible en surface du circuit ne sont pas encore suffisamment mûres pour être industrialisées. Aussi, dans l'état actuel de la technique, le nombre et la complexité des architectures réalisables sont très vite limités car tout ajout de composants électroniques au niveau du pixel se traduit par une diminution relative de la surface photosensible et par une augmentation de la taille générale du capteur.

D'autre part, nous devons valider la technique d'estimation du mouvement en vue d'une intégration par traitement au niveau pixel. C'est-à-dire que nous devons considérer notamment les bruits et imperfections associés. En effet, de nombreuses méthodes existent pour une approche de traitement sur les données vidéo numériques, c'est-à-dire en post-traitement, mais sont-elles adaptées à notre problème ? La référence en termes de performance de perception du mouvement reste encore aujourd'hui celle des êtres vivants, il est donc important que nous considérions aussi ces architectures et modèles de perceptions issus des neurosciences.

Enfin, du point de vue de l'approche scientifique, nous avons adopté une démarche descendante, qui a consisté à commencer par valider de manière algorithmique la stratégie d'estimation du mouvement global, pour progresser ensuite vers son intégration silicium. Nous avons pour cela considéré dans un premier temps nos traitements dans le cas idéal d'images et de vidéos non bruitées, puis nous avons introduit les différents bruits présents au niveau pixel, ce qui nous a parfois conduit à modifier le traitement d'images initial.

Pour mener à bien ce travail, nous avons commencé par un état de l'art sur les capteurs d'images et rétines CMOS, indispensable pour s'imprégner de leurs architectures et des contraintes de conception qui leurs sont associées. A la suite de cela, nous nous sommes intéressés à la stabilisation vidéo et à l'estimation du mouvement, ainsi qu'aux détecteurs du mouvement existants. Nous avons décrit ces deux états de l'art respectivement dans les chapitres I et II de ce manuscrit.

Nous avons ensuite développé une technique d'estimation du mouvement global inter images adaptée à nos spécifications et contraintes matérielles. Cette technique consiste à extraire le mouvement global à partir de mesures locales du mouvement en périphérie de la zone d'acquisition d'images. Elle s'est avérée originale et performante, comme nous le présentons au cours du troisième chapitre.

Enfin le dernier chapitre est consacré à l'intégration sur silicium de la technique à un imageur. Nous établissons alors l'adéquation de l'algorithme développé à l'architecture, et nous explicitons la conception des éléments dédiés à la mesure du mouvement dans le plan focal.

Chapitre I.

CAPTEURS D'IMAGES CMOS,

ETAT DE L'ART

INTRODUCTION

Une image numérique est une matrice de valeurs à deux dimensions représentant l'échantillonnage spatial de l'intensité lumineuse de la scène. En traitement d'image, le terme « pixel » définit un élément de cette image.

Ces images sont obtenues à l'aide d'un capteur d'images, ou imageur. Comme l'illustre la Figure I.1 ci-dessous, ce dispositif est constitué d'éléments optiques projetant la lumière de la scène sur une surface photosensible. Cette surface est en réalité une matrice d'éléments identiques, chacun constitué de composants électroniques et photosensibles, permettant de mesurer la quantité de lumière reçue. L'ensemble des informations de ces éléments formera finalement une image. C'est pourquoi, par extension, nous emploierons le terme « pixel » pour désigner cet élément de l'imageur.

Des circuits complémentaires sont aussi mis en oeuvre pour assurer la lecture des pixels, numériser les signaux acquis et les mémoriser.

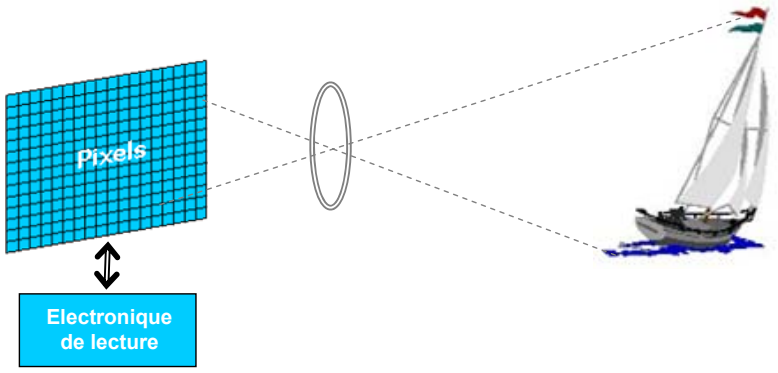


Figure I.1. Architecture générale d'un capteur d'image.

Nous introduisons dans ce chapitre ces éléments constitutifs d'un capteur d'images en technologie CMOS. Nous verrons que la conception de ces dispositifs fait aussi appel à des notions d'optique.

La première partie de ce chapitre sera consacrée à l'élément de base du capteur d'images, et plus précisément du pixel : le photodétecteur. Nous décrirons ensuite quelques éléments d'optique et de photonique. Ces considérations sont nécessaires pour réaliser correctement la projection et l'échantillonnage spatial de la scène lumineuse sur le silicium, mais aussi pour préciser les grandeurs physiques mises en jeu. Enfin, nous aborderons l'électronique de conditionnement et de traitement de cette information électrique photogénérée, pour des applications en imagerie en un premier lieu (les imageurs) et ensuite en vision (les rétines).

I. PHOTODETECTION EN TECHNOLOGIE CMOS

Notre objectif est de concevoir un capteur visuel. Sa fonction élémentaire est donc de mesurer la lumière reçue en la convertissant en information électrique : c'est un composant optoélectronique de type phototransducteur.

I.1. De la lumière à l'électron, phototransduction et photodétection

La photosensibilité d'un semi-conducteur est liée à la capacité des photons, en donnant leur énergie à un atome du réseau cristallin, à amener un électron de la bande de valence à la bande de conduction, créant ainsi une paire électron-trou. Le gap, c'est-à-dire la différence de niveau d'énergie entre la bande de valence et la bande de conduction, fixe l'énergie minimale que doit avoir un photon pour pouvoir créer une paire électron-trou. Dans le cas du silicium, le gap, E_g, est égal à 1,12 eV. Ce matériau est uniquement sensible aux photons de longueur d'onde inférieure à λ_c (Eq. I.1.).

Eq. I.1.
$$\lambda \leq \frac{h.c}{E_g} = \frac{1.24.10^{-6}}{1.12} = 1.11\,\mu m = \lambda_c \text{ (proche infrarouge)}$$

Où « h » est la constante de Planck (6.62×10^{-34} m².kg/s) et « c » la vitesse de la lumière (3×10^{8} m/s).

Un photoconducteur, réalisé à partir d'un barreau de semi-conducteur intrinsèque aux extrémités duquel on a placé deux contacts ohmiques, exploite donc le premier effet des porteurs photogénérés qui est d'augmenter la conductivité du matériau. Une analyse détaillée du fonctionnement de ce dispositif, que l'on pourra trouver par exemple dans [Sze-81], montre que sa sensibilité est proportionnelle au champ électrique dans le matériau.

Or, les jonctions, qu'elles soient métallurgiques ou induites (capacités MOS par exemple), sont le lieu d'un champ électrique important qui permet de séparer les paires électron-trou crées. Il s'ensuit qu'elles sont aussi des photodétecteurs, capables de fonctionner en régime d'accumulation de charges, ce qui n'était pas le cas du photoconducteur.

Le gap du matériau nous a fourni une limite basse, dans le proche infrarouge, à l'énergie des photons capables de créer une paire électron-trou. La profondeur de pénétration associée à la position de la zone de collection va nous donner une limite haute qui se situe à la frontière du visible et de l'ultraviolet.

En effet, l'absorption des photons ne se fait pas à une profondeur fixe : c'est un phénomène qui ne peut être décrit que de manière statistique. On montre ainsi que, pour un photon pénétrant dans le matériau, la probabilité, $p_{abs}(x)$, d'être absorbé avant d'avoir parcouru une distance *x* est :

Eq. I.2.
$$p_{abs}(x) = 1 - e^{-\alpha x}$$

Autrement dit, le flux lumineux décroît de manière exponentielle à partir de la surface et plus de 90% des photons incidents sont absorbés avant d'avoir atteint la profondeur 3/α, «α » étant le coefficient

d'absorption dépendant de la longueur d'onde. Le tableau, ci-dessous, est calculé d'après [Sze-81] et montre l'évolution de 3/α en fonction de la longueur d'onde, à la température de 300 K [2].

λ	$1/\lambda$	α	$3/\alpha$	$Log_{10}(\alpha)$	Couleur
350 nm	0.0028 nm^{-1}	2.4 10^7 m^{-1}	0.125 µm	7.38	Proche ultraviolet
400 nm	0.0025 nm^{-1}	7.4 10^6 m^{-1}	0.405 µm	6.87	Violet extrême
600 nm	0.0016 nm^{-1}	4.7 10^5 m^{-1}	6.38 µm	5.67	Orangé moyen
800 nm	0.00125 nm^{-1}	1.2 10^5 m^{-1}	25 µm	5.08	Proche infrarouge

Tableau I.1. Coefficient d'absorption dans le silicium en fonction de la longueur d'onde.

A partir de ces valeurs, il est possible de construire, par régression linéaire, une expression équivalente pour une longueur d'onde comprise entre 400 nm et 800 nm :

Eq. I.3. $$\log_{10}\alpha \approx 3{,}27 + 1440 \times \frac{1}{\lambda_{nm}}$$

Il apparaît clairement que, pour qu'un phototransducteur en silicium présente une sensibilité raisonnable dans le proche ultraviolet (350 nm), il doit collecter les photocharges à moins de 150 nm de sa surface.

La quantité de charges électriques générées est donc proportionnelle au nombre de photons pénétrant le silicium par unité de surface et de temps. Ce rapport du nombre d'électrons photogénérés et collectés par le nombre de photons incidents définit le **rendement quantique** (Eq. I.4), il représente un critère essentiel de performance d'un photodétecteur et il est noté « η ».

Eq. I.4. $$\eta = \frac{\text{Nb d'électrons photogénérés}}{\text{Nb de photons incidents}}$$

Un autre paramètre utile pour caractériser le photorécepteur est son **rendement de détection (ou réponse spectrale)**. Il exprime la relation de proportionnalité entre la quantité de charges générées par unité de temps et la puissance lumineuse reçue. Il est lié au rendement quantique par :

Eq. I.5. $$R = \frac{I}{P} = \frac{q.\lambda.\eta}{h.c} \approx 0.805.\lambda_{\mu m}.\eta \quad (\text{Amp./Watt.})$$

Pour un rendement quantique de un (maximum) et pour une longueur d'onde comprise entre 400nm et 700nm (visible), on a donc un courant photogénéré respectivement compris entre **320 mA/W et 560 mA/W**.

[2] *Nous avons inclus en page Annexe B la courbe complète des valeurs de α en fonction de la longueur d'onde.*

I.2. Photodétecteurs

Le photodétecteur constitue le premier élément de la chaîne de traitement du signal d'un capteur d'image. Dans un procédé de fabrication CMOS standard, nous disposons de trois familles de composants photosensibles : les photodiodes, les phototransistors et les photogrilles.

I.2.a. Photodiode

La photodiode est aujourd'hui le photodétecteur standard dans les capteurs d'images. La zone de collection est formée par la zone de charge d'espace de la jonction polarisée en inverse.

Étant donnée l'architecture d'un pixel à base de photodiode, deux paramètres sont essentiels pour qu'un pixel soit performant en imagerie : son rendement de détection et sa capacité de jonction.

- Photodiodes CMOS

Remarque : *Les dimensions que nous indiquons ici correspondent à une technologie 0,35 µm et peuvent être considérées comme typiques. Sauf indication contraire, la taille de la zone de charge d'espace de la jonction est donnée pour une tension inverse de 3,3 V*.

La Figure I.2 résume la structure d'un circuit intégré CMOS sur substrat P. On peut y identifier quatre types de jonction :

1. N+/P-Well : une zone de charge d'espace d'environ 250 nm placée à 200 nm de l'interface oxyde-substrat. Une jonction présentant un maximum de sensibilité dans le vert ou le jaune.
2. P+/N-Well : une zone de charge d'espace d'environ 150 nm placée à 200 nm de l'interface oxyde-substrat. Une sensibilité chromatique équivalente à la précédente, avec un rendement quantique un peu moins élevé.
3. N-Well/P-Well : une zone de charge d'espace d'environ 1,5 µm recevant les photons « par la tranche » et profonde d'environ 3 µm. Une diode présentant « forcément » un bon rendement quantique et une sensibilité chromatique « large ».
4. N-Well/P-Sub : une zone de charge d'espace d'environ 3 µm placée à 3,5 µm de l'interface oxyde-substrat. Une diode présentant un bon rendement quantique avec une sensibilité chromatique plutôt dans le rouge.

4'. N-Well/P-Sub : une zone de charge d'espace d'environ 1.5 µm recevant les photons « par la tranche » et profonde d'environ 3 µm. Une diode présentant un très bon rendement quantique avec une sensibilité chromatique « large ».

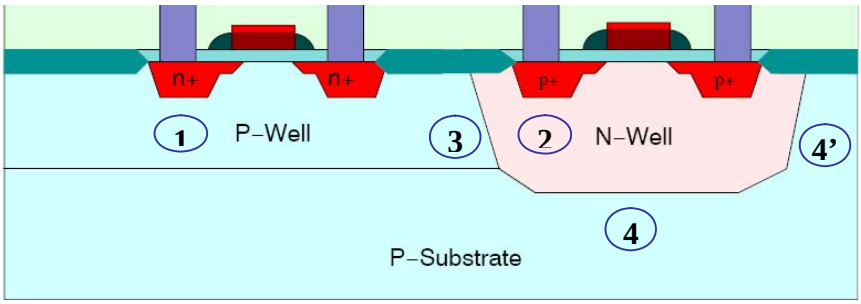


Figure I.2. Vue en coupe d'un circuit CMOS standard (technologie double puits sur substrat P--)

Le rendement de détection dépend donc fortement de la profondeur où se situe la zone de collection (cf. Figure I.5). Ainsi, une diode N+/Psub sera plus sensible aux longueurs d'onde proches du bleu qu'une diode Nwell/Psub qui captera préférentiellement la lumière rouge [Lin et al.-02].

En revanche, les photocharges ne sont pas seulement collectées dans la zone de charge d'espace de la jonction. En effet, des électrons (resp. des trous) photogénérés dans l'anode (resp. la cathode) peuvent gagner, par diffusion, la zone de charge d'espace où ils sont accélérés puis collectés.

Cette diffusion latérale a été exploitée par [Dierickx et al.-97] qui réussit à collecter une grande partie des photocharges créées ailleurs que dans la zone de collection (cf. Figure I.3). De cette manière une grande partie de la surface du pixel est utilisée, alors qu'elle n'est que de 40 % environ dans les pixels classiques. La courbe de potentiel de la section « A » (cf. Figure I.3) montre bien une « cuvette » de potentiel dans la zone substrat dopée « p- », ce qui a tendance à maintenir les photocharges dans ce volume jusqu'à ce qu'ils atteignent la zone de collection.

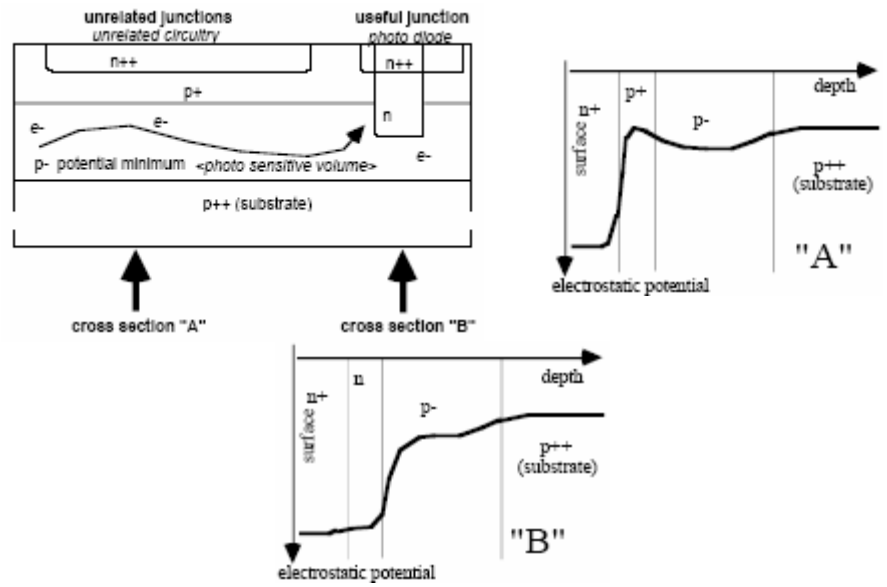


Figure I.3.Constitution du pixel proposé par [Dierickx et al.-97] avec les courbes de potentiel caractéristiques, associées aux sections A et B.

Le courant d'obscurité, provenant des courants de fuite dans le photodétecteur en l'absence de tout flux lumineux incident, est de l'ordre de 30 nA/cm² et peut être réduit jusqu'à une dizaine de nA/cm² en adoptant quelques précautions lors du dessin des masques, notamment au niveau des jonctions des composants [Wu et al.-04] [Tian et al.-01].

→ Double jonction Pdiff/Nwell et Nwel/Psub.

La configuration double jonction, associant la réponse spectrale des deux photodiodes (cf. Figure I.4) pour réduire le nombre de filtres optiques colorés nécessaires à la reconstitution d'une image couleur, a été mise à profit par [Lu-01] et [Findlater et al.-03a]. La séparation spectrale est moins satisfaisante que dans

le cas de l'approche classique basée sur une mosaïque de Bayer[3]. En revanche, le nombre d'artefact est réduit.

En ce qui concerne la photodétection, le pic de la réponse spectrale de la photodiode Pdiff/Nwell se situe à environ 530 nm alors que celui de la jonction Nwell/Psub se trouve autour de 710 nm. Cela est bien en accord avec ce que nous avons évoqué précédemment, à savoir que la profondeur de création des charges libres dépend de la longueur d'onde du photon incident. Effectivement, les photons absorbés dans la photodiode Pdiff/Nwell, dont la zone de déplétion se situe près de la surface du semiconducteur, ont une longueur d'onde plus faible que ceux absorbés par la photodiode Nwell/Psub qui possède une zone de déplétion plus enterrée.

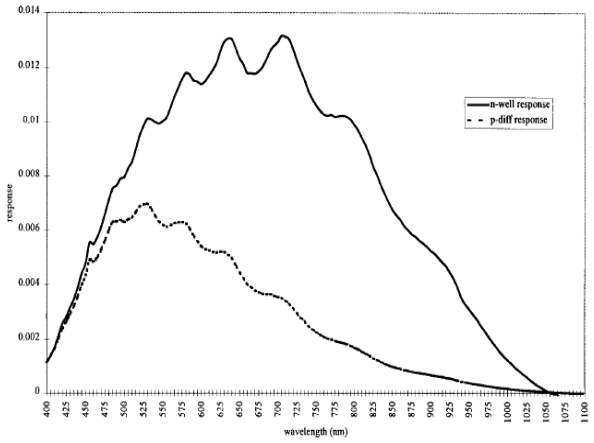


Figure I.4. Réponses spectrales normalisées de deux photodiodes de type Pdiff/Nwell et Nwel/Psub [Simpson et al.-99].

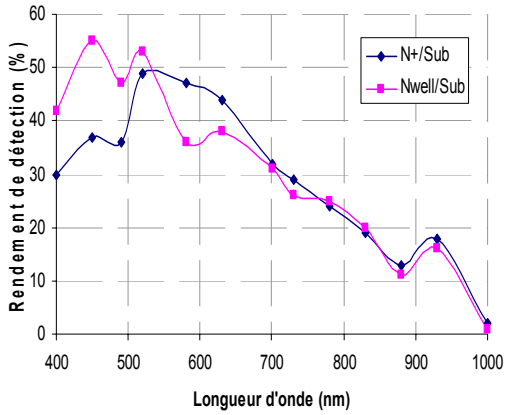


Figure I.5. Rendement de détection typiques des photodétecteurs de types : Ndiff/Psub et Nwell/Psub, en CMOS 250nm [Magnan et al.-04].

- Photodiode verrouillée, ou « pinned »

La photodiode verrouillée est issue de la technologie CCD [Burkey et al.-84]. Il s'agit d'une diffusion N sur un substrat P avec une fine couche de diffusion P+ en surface qui permet de fixer le potentiel de surface à celui du substrat, d'où son appellation « verrouillée ». Ce potentiel est fixé de telle manière que la zone N soit totalement déplétée afin de permettre un transfert total des photocharges (cf. § III.1.a. pour le fonctionnement du pixel associé).

Les photodiodes verrouillées permettent d'obtenir un facteur de détection élevé (cf.

Figure I.6) et un courant d'obscurité plus faible de l'ordre de 40 pA/cm² [Findlater et al.-03b], au prix de modifications de quelques niveaux du procédé de fabrication CMOS.

---

[3] Le motif de Bayer est employé pour déterminer par interpolation la couleur de la lumière reçue par un pixel. Il se matérialise sur la matrice de pixels par l'alternance de deux lignes de pixels. L'une est constituée d'une succession de filtre vert, puis rouge, puis vert à nouveau, etc… ; et l'autre d'une succession de filtre bleu, puis vert, puis bleu, etc…

Le premier capteur d'image utilisant une photodiode verrouillée a été décrit par [Lee et al.-95]. Leur objectif était de combiner les performances des photodétecteurs issus de la technologie CCD avec les avantages en termes d'intégration système de la technologie CMOS. Aujourd'hui, les photodiodes verrouillées constituent le standard dans les capteurs d'images car l'architecture des pixels qui les utilisent (à 4 transistors, cf. § III.1.a.) permet de réduire drastiquement la taille des pixels en partageant certains transistors, ainsi que de diminuer le bruit de lecture [Findlater et al.-03b] [Kasano et al.-05].

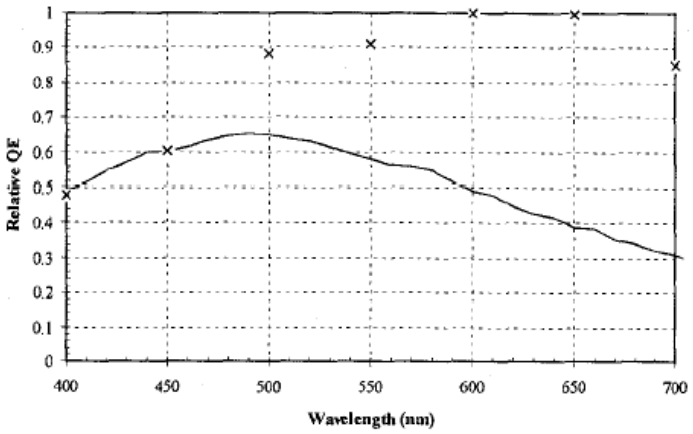


Figure I.6. Facteur de détection relatif d'une photodiode verrouillée (croix) et celle d'un capteur CCD interligne (courbe) [Guedash et al.-97].

- Photodiode en silicium amorphe hydrogéné, a-Si :H.

Il y a une dizaine d'années, on pensait que la taille des pixels ne pourrait pas descendre au dessous de 5×5 μm² environ à cause des limites optiques [Fossum-97]. Elle est actuellement de l'ordre de 3.5×3.5 μm² dans les produits vendus et de 2.5×2.5 μm² pour la nouvelle génération en cours de développement. Cette diminution a été rendue possible à la fois grâce à des progrès optiques et électroniques. Cependant, la focalisation optique est de plus en plus difficile et coûteuse à réaliser, à tel point qu'une limite en l'état actuel des technologies est prédite autour de 2×2 μm² [Rhodes-04].

Une alternative apparue au début des années 90 [Fisher et al.-92] semble prometteuse et est actuellement en cours de développement chez certains fabricants de semi-conducteurs tel que Agilent [Theil-03]. Elle consiste à intégrer les photodétecteurs au dessus des différents niveaux d'oxydes, de métaux et de polysilicium constituant les circuits CMOS (cf. Figure I.7). Cette technologie est aussi appelée « à dépôt fin sur ASIC » ou « à photodétecteurs surélevés »[4]. De plus, cette dissociation de la fabrication de l'élément photosensible du reste du circuit est très intéressante car elle permet de concevoir non seulement des photodétecteurs mais aussi des photoémetteurs de type OLEDs[5] [Theil-03]. Cela ouvre donc la voie à de nouvelles et nombreuses applications [Dong et al.-05] [Karim et al.-03] [Benthien et al.-00].

[4] *« Thin Film on Asic » (TFA) ou « elevated photodetectors » en langue anglo-saxonne.*

[5] *« Organic Light-Emitting Devices ».*

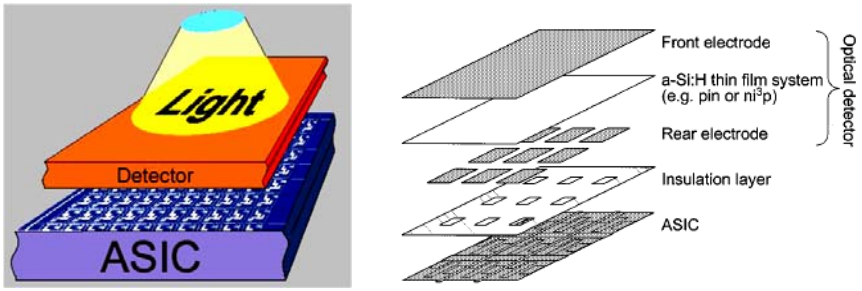


Figure I.7. Architecture de photodétecteur de nouvelle génération, en procédé TFA.

Le procédé de fabrication des photodiodes en silicium amorphe nécessite trois couches supplémentaires (cf. Figure I.7) dont le dépôt n'était pas bien maîtrisé à ses début mais qui le devient désormais [Theil-03] [Lule et al.-00a]. La couche supérieure est un oxyde conducteur transparent, la couche intermédiaire réalise une jonction de type pin[6] et la couche inférieure constitue l'électrode arrière.

Cette approche dispose de plusieurs atouts. Tout d'abord, le gap du silicium amorphe étant de 1.7 eV, sa réponse spectrale est alors concentrée sur le domaine visible (cf. Figure I.8). De plus, le rendement de détection est élevé (cf. Figure I.8), il peut atteindre 80%. Un autre élément important est que cette réponse spectrale peut être ajustée en fonction de l'épaisseur de la couche « i » de la diode pin (cf. Figure I.8 droite), mais aussi en fonction de la tension inverse appliquée à la photodiode [Zhu et al.-95]. Le courant d'obscurité de ces structures est de l'ordre de 100 pA/cm² [Lule et al.-00a].

Cependant, des problèmes subsistent, notamment un vieillissement non maîtrisé des couplages résistifs entre pixels voisins, et une non uniformité spatiale importante à cause des disparités des rendements de détection entre pixels voisins[7].

I.2.b. Phototransistor

Nous pouvons également réaliser, en procédé CMOS standard, un détecteur de type phototransistor. Il s'agit d'un transistor bipolaire dont le courant de base est le photocourant généré par la jonction base-émetteur qui est polarisée en inverse. Le photocourant résultant de l'effet transistor est alors amplifié par le gain β du transistor, de même que le courant d'obscurité qui provient de la jonction base-émetteur et qui a donc les caractéristiques discutées précédemment. La consommation qu'il engendre est elle aussi supérieure à celle d'une photodiode d'un facteur β. De plus, sa capacité d'entrée est grande ce qui en fait un photodétecteur « lent ».

---

[6] Une diode PIN est formée par la superposition un volume dopé N, une zone neutre I et une zone dopée P. La zone neutre permet d'agrandir la zone de collection.

[7] Encore appelé « PRNU » pour « PhotoResponse NonUniformity ».

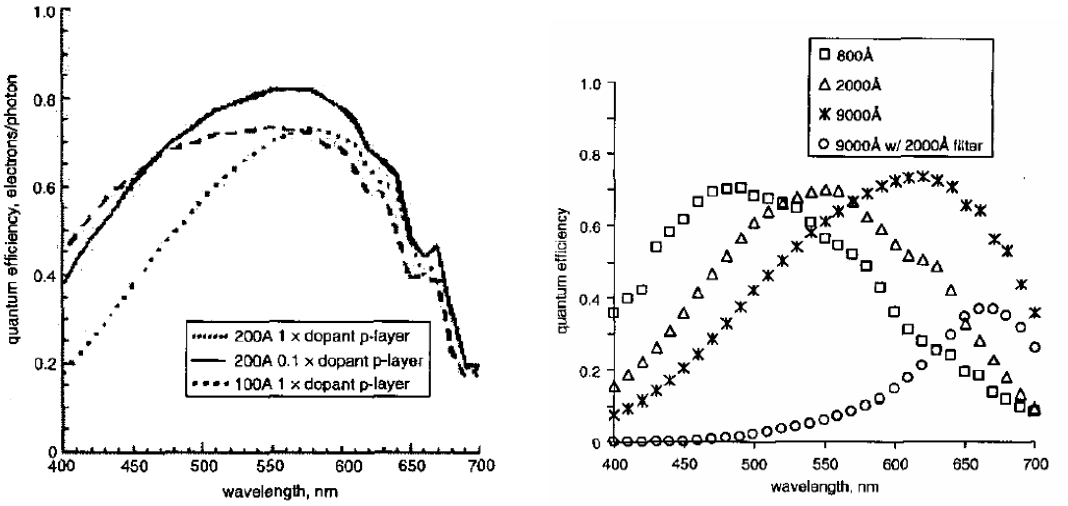


Figure I.8. Rendement de détection de diodes a-Si:H de type pin [Theil -03]. Courbes de gauche : en fonction de l'épaisseur de la couche « P » de la diode, Courbes de droite : en fonction de l'épaisseur de la couche intermédiaire « i ».

I.2.c. Photogrille

Le détecteur photogrille a été proposé par [Mendis et al.-93] afin d'obtenir une structure de pixel qui permet une réduction du bruit de lecture plus efficace (cf. § III.1.a. page 23) et ainsi une meilleure qualité d'images. Les photocharges créées sont piégées dans la zone de collection présente sous la photogrille qui est portée à un potentiel élevé. Pendant l'intégration, le transistor TX est bloqué et la diode de lecture (à droite de TX sur la Figure I.9) est préchargée. Au moment de la lecture, TX est rendu passant et PG est porté au potentiel du substrat. La valeur de la capacité MOS s'annule et toutes les charges qu'elle contenait sont transférées dans le drain et la source de TX. Il s'ensuit une diminution de la tension aux bornes de la diode de lecture.

Le conditionnement est alors à quatre transistors et possède les mêmes avantages que dans le cas de la photodiode verrouillée, où la totalité des charges peut être transférée de la photogrille sur une capacité équivalente plus petite pour atteindre une grande sensibilité.

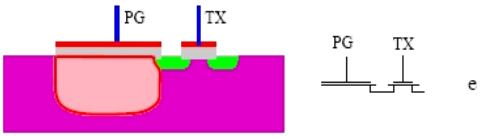


Figure I.9. Structure du détecteur photogrille.

Le courant d'obscurité est quant à lui de l'ordre de plusieurs dizaines de nA/cm² [Tian et al.-01], mais l'inconvénient majeur du détecteur photogrille est la présence du polysilicium de grille qui réduit grandement le rendement de détection dans le bleu, comme nous l'illustrons sur les courbes de la Figure I.10.

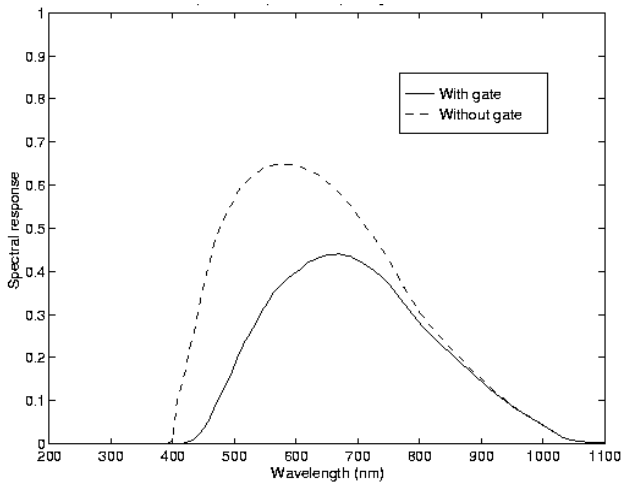


Figure I.10. Réponse spectrale typique pour un détecteur photogrille [Moini-99].

Dans cette première partie, nous avons considéré la réponse spectrale des divers photodétecteurs seuls. Cependant, les capteurs d'images qui sont dédiés à l'acquisition et la restitution des images sont composés d'éléments optiques tels que des filtres colorés et des microlentilles afin de restituer des images couleurs et augmenter la sensibilité. Ces éléments influent sur les performances finales des capteurs et surtout sur le dimensionnement des photodétecteurs et de leur électronique de conditionnement. Il est donc important de les considérer afin d'être capable de concevoir correctement cette électronique.

II. OPTIQUE

Les capteurs d'images ou rétines sont des systèmes électroniques conçus pour percevoir et quantifier un flux photonique incident. L'instrumentation optique est une science à part entière, nous ne présentons ici que ce qui nous paraît essentiel en vue de concevoir un capteur d'image ou une rétine électronique.

Nous abordons notamment les bases théoriques de la mise au point des éléments optiques employés dans un capteur CMOS afin d'être capables de mener les mesures expérimentales dans des conditions adéquates. Nous définissons l'expression des grandeurs et unités à la base des caractérisations photoniques, et enfin nous présentons le chemin optique dans un imageur CMOS. L'objectif poursuivi est de déterminer la quantité de charges électriques effectivement photogénérées, en prenant en considération les différents éléments optiques présents dans un capteur d'image et qui jouent un rôle important dans la perception lumineuse. Pour plus de détails, nous conseillons de se référer à [Ryer-97].

II.1. Échantillonnage spatial

Tout capteur d'image réalise un échantillonnage spatial de la scène grâce à sa matrice de pixels.

II.1.a. Tesselations

Dans le cas d'un imageur, c'est-à-dire un capteur d'images dédié à l'imagerie, la tessellation des pixels devra prédisposer le capteur à restituer l'image la plus nette possible pour une visualisation ultérieure. Pour arriver à ces fins, des critères et métriques de perception humaine ont été établis afin d'évaluer cette netteté de manière non subjective [Bovik-00]. Des études théoriques [Dubois-85] et expérimentales [Moini-

99][Tirunelveli et al.-02] ont été menées dans le cas de signaux multidimensionnels, en considérant en premier chef l'application vidéo. Ces derniers travaux comparent les deux plus intéressantes tessellations : hexagonale et rectangulaire (cf. Figure I.11), pour conclure que l'organisation rectangulaire donne d'aussi bons résultats que l'hexagonale. Par contre, en considérant les contraintes technologiques de fabrication des circuits CMOS, la répartition la mieux adaptée est la répartition rectangulaire, particulièrement symétrique et donc plus simplement intégrable. C'est pourquoi tous les imageurs CMOS sont de nos jours conçus ainsi.

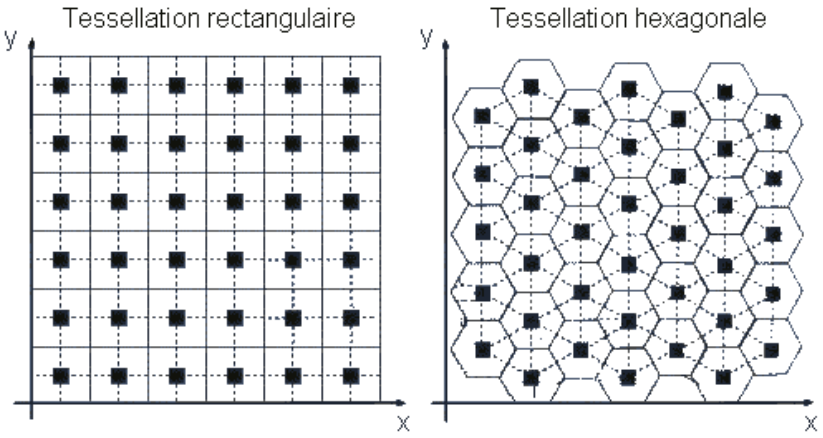


Figure I.11. Tessellation rectangulaire et hexagonale.

II.1.b. Interdépendance entre pixels et optique : l'aliasing

Comme pour tout échantillonnage, il y a périodisation de la réponse fréquentielle. Si le spectre de la scène à acquérir est plus large que la moitié de la fréquence d'échantillonnage, alors le problème de « **l'aliasing »** ou du **repliement** apparaît [Lyons-01]. Il se matérialise sur les images par des structures appelées « moirés » résultant de la superposition fréquentielle.

La Figure I.12 suivante illustre le repliement de fréquence pour un signal monodimensionnel de fréquence centrale f_0 et de bande spectrale Δf, la partie hachurée est la zone de recouvrement des spectres :

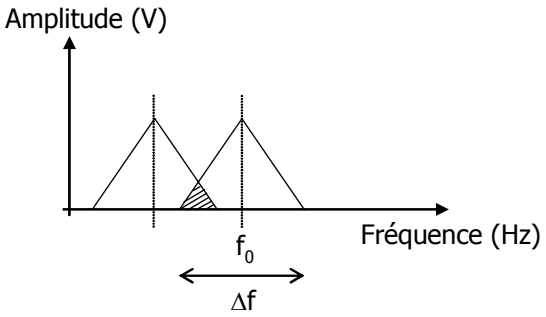


Figure I.12. Repliement de spectre.

Les caractéristiques et la mise au point de l'optique doivent donc être déterminées en fonction de la taille des pixels de l'imageur, ou inversement.

Pour cela, considérons tout d'abord un objectif constitué d'une lentille unique. Il est caractérisé de manière unique par sa **distance focale f** et par son **nombre d'ouverture n** = *f*/*D*, D étant le diamètre de la lentille. La Figure I.13 rappelle la construction de l'image réelle d'un objet placé au-delà du foyer. F et F' sont les foyers, respectivement objet et image, de la lentille. Ω est l'angle solide sous lequel est vue la lentille depuis l'objet.

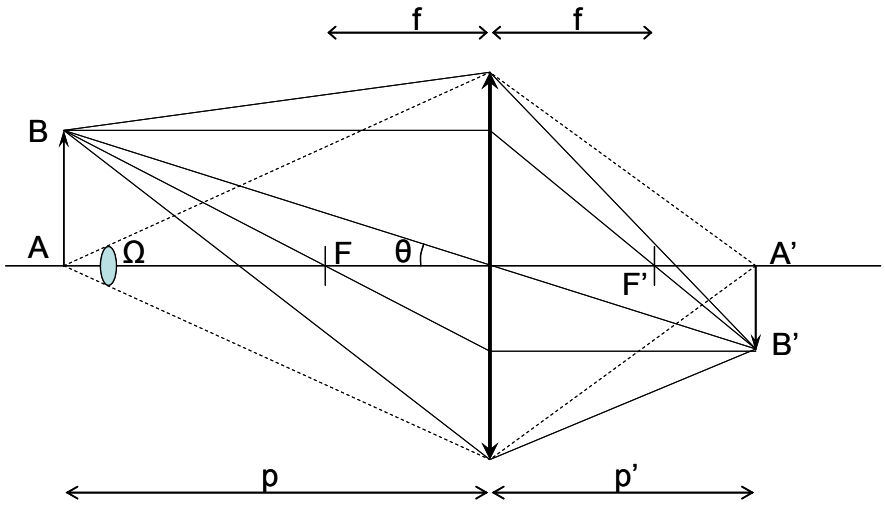


Figure I.13 : Image réelle d'un objet créée par une lentille

En appliquant le théorème de Thalès aux différents triangles de la construction, on obtient la relation fondamentale :

$$\frac{1}{p}+\frac{1}{p'}=\frac{1}{f}$$

La position de l'image et sa dimension se déduisent de la même façon :

$$p'=\frac{p\ f}{p-f}\approx f \quad \text{et} \quad A'B'=\frac{p'}{p}AB=\frac{f}{p-f}AB\approx f\frac{AB}{p}\approx f\,\theta$$

Les approximations correspondent au cas où *p* >> *f*. Dans ce cas, le rapport AB/p représente la tangente de l'angle θ sous lequel est vu l'objet depuis la lentille (rappelons que si *AB* << *p*, tan θ ≈ θ).

Classiquement, on considère que la distance focale de l'objectif doit être du même ordre que la diagonale de l'imageur. Prenons l'exemple d'un imageur 800×600 caractérisé par une diagonale de 5mm. Pour limiter les défauts, observons une petite marge de sécurité et considérons une lentille de 8mm de focale.

De toute rigueur, l'imageur doit être placé à la distance p' de la lentille pour recevoir une image nette de l'objet. La distance p étant susceptible de changer, il faudrait pouvoir modifier la distance imageur-lentille en permanence. Nous allons cependant voir qu'il est possible de placer l'imageur à une distance fixe.

Considérons le point A', image du point A. Lorsque $p \to \infty$, A' se confond avec F'. Supposons que l'imageur soit placé à une distance p'_i de la lentille, avec $p'_i > f$. Sur l'imageur, l'image du point A n'est plus un point, mais un disque de diamètre d' tel que :

$$\frac{d'}{D} = \frac{p'_i - f}{f} \qquad \text{Eq. I.6.}$$

Expression dans laquelle D est le diamètre de la lentille. Si le point A se rapproche, il vient un moment où son image A' est exactement dans le plan de l'imageur. Si A dépasse cette position, son image sur l'imageur est à nouveau un disque de diamètre d' vérifiant la relation :

$$\frac{d'}{D} = \frac{p' - p'_i}{p'} \qquad \text{Eq. I.7.}$$

Ce « défaut de mise au point » reste imperceptible tant que d' reste plus petit que la taille d'un pixel. En fixant une valeur minimale pour p, on peut donc en déduire le nombre d'ouverture de la lentille à utiliser. En considérant que l'on a le même d' dans les deux équations précédentes, on obtient :

$$n = \frac{f}{D} = \frac{f^2}{d'(2p - f)}$$

$$p'_i = f + n\,d'$$

Pour un imageur de taille de pixels de 5x5 µm², si l'on veut une profondeur de champ s'étendant de 1m à l'infini, il faudra choisir un objectif de nombre d'ouverture n > 6,43. La valeur normalisée immédiatement supérieure est n = 8. On obtient alors p'_i = 8,04 mm et une profondeur de champ s'étendant de p = 80 cm à l'infini.

II.2. Chemin optique d'un imageur

La quantité de lumière reçue par la zone photosensible d'un capteur d'image est loin d'être semblable à la quantité émise par la source lumineuse. En effet cette lumière reçue sur le silicium provient de la réflexion des objets de la scène et du passage par plusieurs milieux homogènes d'indices de réfraction distincts.

Comme illustré sur la Figure I.14, ces éléments optiques que la lumière doit traverser dans un imageur CMOS sont : une couche protectrice, une lentille, et un filtre de couleur (rouge, vert, ou bleu pour constituer le motif de Bayer).

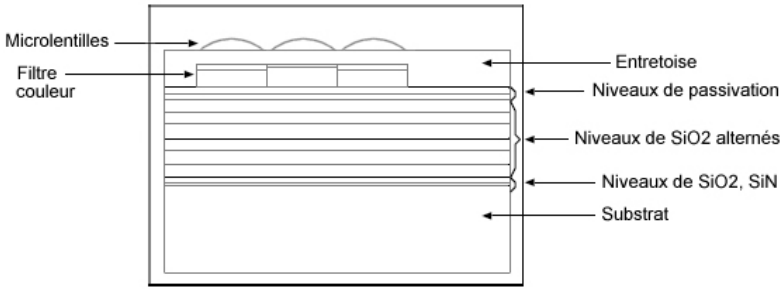


Figure I.14. Empilement des couches optiques sur le substrat silicium [Vaillant & Hirigoyen-04].

Les expériences menées chez STMicroelectronics indiquent que seulement 1% de la lumière atteint le silicium. On considèrera alors, par la suite, un coefficient de transmission optique de 1%.

II.3. Photométrie et radiométrie

Nous avons présenté les photodétecteurs CMOS, nous avons ensuite déterminé en première approximation l'influence de l'optique dans les imageurs, nous décrivons maintenant les éléments essentiels de photométrie et radiométrie qui permettront de caractériser les performances de tout capteur d'images. Ces performances sont effectivement fonction du nombre de photons incidents, qu'il nous faut savoir quantifier.

Nous présentons également dans cette section le calcul, incontournable en imagerie, du nombre d'électrons effectivement photogénérés et qui constituent l'information lumineuse finalement perçue par le pixel.

II.3.a. Flux photonique, puissance, illumination

Il est impératif lors de toute caractérisation photonique de connaître précisément la quantité de photons incidents. Deux paramètres sont couramment employés pour décrire les conditions expérimentales d'éclairement : l'illumination et la puissance lumineuse. **L'illumination** tout d'abord, s'exprime en **lux**. Pour des raisons pratiques, cette unité lie l'énergie lumineuse par unité de surface à la perception de l'œil humain. Ainsi un objet situé à 30 cm d'une bougie représente une illumination de 10 lux. La **puissance lumineuse** est quant à elle exprimée en **W.m^{-2}**.

Cependant, l'intérêt de ces deux unités est de quantifier le **flux photonique** incident, que l'on exprime en **photons.m^{-2}.s^{-1}**. Nous exprimons ci-dessous les relations qui existent entre ces trois paramètres.

Le point de départ est l'énergie du photon, qui est définie par :

Eq. I.8. $$\mathrm{Ep} = \mathrm{h} \times \nu = \frac{\mathrm{h} \times \mathrm{c}}{\lambda} \quad (J)$$

Où « ν » est la fréquence du photon (Hz).

La puissance lumineuse est définie par :

Eq. I.9. $$Flux\,photons = \frac{Puiss.\,lum.}{Ep}$$

L'illumination et la puissance lumineuse sont des paramètres qui dépendent de la longueur d'onde de la radiation. On se place alors à une longueur d'onde donnée, par exemple à 555 nm qui correspond au pic de meilleure perception d'un œil humain [CIE-04], et dans ce cas l'énergie d'un photon devient :

Eq. I.10. $$\mathrm{Ep}_{555} = \frac{3.10^{8} \times 6.62.10^{-34}}{555.10^{-9}} = 3.58.10^{-19}\ \mathrm{J}$$

D'où le flux de photons par watt :

Eq. I.11. $$1\,\mathrm{W} = \frac{1}{3.58\,.10^{-19}} = 2.79\,.10^{18}\ \mathrm{photons.s^{-1}}$$ à 555 nm

De plus, à 555 nm, on a [Ryer-97] :

Eq. I.12. $$1\,\mathrm{W.m^{-2}} = 683\,\mathrm{lux} \quad \text{ie.} \quad 1\,\mathrm{lux} = 1.46\,.10^{-3}\ \mathrm{W.m^{-2}}$$

On obtient alors pour une illumination de 1 lux le flux de photons suivant :

Eq. I.13. $$1\,\mathrm{lux} = \frac{1.46\,.10^{-3}}{3.58\,.10^{-19}} = 4.09\,.10^{15}\ \mathrm{photons.m^{-2}.s^{-1}}$$ à 555 nm

II.3.b. Charges électriques photogénérées, « utiles »

Nous sommes maintenant capables de déterminer le flux de photons incidents sur une surface donnée de silicium, voyons comment nous obtenons le nombre d'électrons collectés.

Mathématiquement, le nombre d'électrons collectés par un pixel s'écrit :

Eq. I.14. $$I = T.\int_x \int_y \int_\psi E(x,y,\lambda).Sr(x,y).\eta(\lambda).dx.dy.d\lambda$$

Où (x,y) représente les coordonnées continues sur le plan focal du capteur, « ψ » est le spectre de longueurs d'ondes « λ » visibles, « $\eta(\lambda)$ » est le rendement de détection, « T » est le temps d'exposition du pixel à la lumière incidente, « $E(x,y,\lambda)$ » est le flux lumineux incident sur le point de coordonnées (x,y) du pixel[8] et « $S_r(x,y)$ » est la réponse spatiale du pixel.

Si nous nous plaçons maintenant à une longueur d'onde de 555nm et que nous considérons que la réponse spatiale du pixel est de « 1 » sur sa surface photosensible. Alors le nombre d'électrons photogénérés se ramène à l'expression :

Eq. I.15. $$\mathrm{Nb.Electrons}_{555} = \mathrm{T} \times \mathrm{K} \times \mathrm{S} \times \eta(555)$$

Où « K » est le flux de photons reçus, puis « S » est la surface photosensible du pixel considéré (lumière concentrée sur le photodétecteur).

A titre d'exemple, en faisant l'hypothèse d'un coefficient de transmission optique de 1%, d'une longueur d'onde de 555 nm, et en prenant le cas d'une photodiode « Nwell/Psub » recevant une quantité de lumière

[8] *Incluant le facteur de transmission résultant du conditionnement optique de l'imageur (cf. §Chapitre I.II.2.).*

équivalente[9] à une surface de 3×3µm², sous une puissance lumineuse de 1W.m^{-2}, pendant 1ms, et de rendement de détection de 45% (cf. Figure I.5 page 11) la quantité de charges « utiles » photogénérés vaut alors:

$$\text{Nb.Electrons}_{555} = 10^{-3} \times 10^{-2} \times 2.79 \times 10^{18} \times 9 \times 10^{-12} \times 0.45 \sim 115 \; \textit{électrons}$$

Cela signifie que nous devons concevoir une électronique pixel quasiment capable de compter l'électron. Cette valeur nous donne également une information référence à mettre en rapport avec le bruit électronique présent dans le capteur.

III. CAPTEURS POUR L'IMAGERIE, LES IMAGEURS

L'objectif est de restituer la meilleure image au sens de la perception humaine. Quelques exemples de produits commerciaux mettant en oeuvre des imageurs CMOS sont les webcams, les téléphones cellulaires ou assistants numériques personnels avec appareils photos intégrés, ou les jouets. D'autres dispositifs tels que les souris optiques et les lecteurs de codes barres bénéficient aussi de cette technologie et intègrent des imageurs spécifiques à grande vitesse d'acquisition et faible résolution.

On remarque que l'essentiel du marché concerne les dispositifs grand public à bas coût, où les performances requises sont modérées. Cependant les capteurs d'images CMOS font l'objet d'efforts de recherche et de développement très importants. Ceux-ci sont en effet motivés par les fabricants d'imageurs qui souhaitent se démarquer de la concurrence pour augmenter leur part sur le marché très fructueux des imageurs[10]. Ainsi leurs performances s'améliorent continuellement et égalent celles des capteurs CCD de moyenne gamme [Magnan-03]. Il en résulte qu'ils commencent à se destiner à certaines applications qui étaient réservées au CCD auparavant, comme la photographie numérique où les plus grands fournisseurs en vantent les performances [Canon-05].

III.1. Conditionnement de l'information électrique, le pixel image

Nous avons vu comment convertir l'information lumineuse en charges élémentaires, voyons maintenant comment convertir ces photocharges en tension, information électrique que nous pourrons exploiter plus aisément ensuite pour constituer une image.

L'objectif est ici de conditionner l'information électrique engendrée par le photodétecteur afin d'obtenir le meilleur rapport signal sur bruit, dans un encombrement minimum. Diverses structures existent pour cela [Fossum-97], que l'on peut classer en deux familles : les pixels à intégration et les pixels à lecture directe.

---

[9] C'est-à-dire que l'on considère ici l'effet de concentration du flux lumineux par la microlentille sur la photodiode (la surface de la photodiode étant elle-même inférieure à cette surface de 3×3 µm²).

[10] Les ventes mondiales d'imageurs CMOS pour téléphones mobiles se sont en effet élevées à 35.2 millions d'unités en 2004, pour un marché global incluant les imageurs CCD (140 millions d'unités) de 3.4 Milliards de dollars [SST-04] [Elect. Int.-05].

Ces derniers fournissent, en continu, une tension qui est fonction du flux lumineux qu'ils reçoivent. Ils sont souvent logarithmiques[11] et souffrent d'un bruit important qui les pénalise pour des applications d'imagerie [Matou-03]. Aussi nous nous intéressons aux pixels procurant le meilleur compromis coût-performances en l'état actuel des technologies : les pixels à intégration. Ces pixels accumulent les charges photogénérés pendant un temps d'exposition donné puis restituent une tension dépendant de cette quantité de charge.

Nous présentons les paramètres essentiels qui caractérisent un pixel, venant compléter ceux du photodétecteur (réponse spectrale et courant d'obscurité) :

- La taille et le facteur de remplissage.

On compte aujourd'hui des millions de pixels dans un imageur. Leur taille doit par conséquent être minimisée à l'extrême afin de réduire le coût de fabrication du capteur.

D'autre part, un pixel CMOS est constitué d'un photodétecteur et d'une électronique de conditionnement. Le rapport de la surface du photodétecteur à celle du pixel entier définit le **facteur de remplissage** qui renseigne sur la structure et les performances du pixel. En effet, dans le cas général et pour une taille de pixel donnée, plus le facteur de remplissage est petit, plus le rapport signal sur bruit est faible.

- Le gain de conversion « G_{conv} ».

Les photocharges générées dans un pixel actif sont converties en tension dans le pixel. Cette tension contiendra alors l'information de la quantité de lumière reçue par le pixel. Cette conversion charges-tension est réalisée par l'intermédiaire de la capacité équivalente sur la grille du transistor suiveur, que nous appelons capacité de conversion C_{conv}.

Nous définissons alors le gain de conversion, paramètre essentiel qui caractérise l'efficacité du pixel à convertir en tension les charges (électrons) photodétectées, par le rapport suivant :

$$G_{conv} = \frac{q}{C_{conv}} \qquad \text{Eq. I.16.}$$

où « C_{conv} » est la capacité équivalente sur la grille du transistor suiveur de lecture du pixel (cf. Figure I.15 à Figure I.20) et « q » est la charge de l'électron.

La conception du pixel consistera alors à maximiser à la fois ce gain de conversion et le nombre de photocharges collectées.

Ce gain de conversion est particulièrement important pour deux raisons essentielles : il détermine d'une part la sensibilité de l'imageur et il agit d'autre part directement sur le bruit temporel de sortie de l'imageur (cf. § III.2.a. page 29).

[11] *C'est-à-dire dont la caractéristique de la tension de sortie est proportionnelle au logarithme de l'intensité lumineuse incidente.*

III.1.a. Pixels actifs

Ces pixels sont dits actifs car ils disposent d'une amplification en courant du signal photogénéré. Cet étage d'amplification est la plupart du temps obtenu par un transistor suiveur. Le signal utile, image de la quantité de lumière reçue, est obtenu par différence de deux potentiels échantillonnés en sortie du transistor suiveur à deux instants séparés par un intervalle Δt qui est le temps d'exposition. Le premier potentiel est celui de référence, après initialisation du pixel, et le second est fonction de la quantité de charges photodétectées pendant l'intervalle de temps Δt.

- APS-3T

Dans le cas d'un pixel actif à trois transistors, l'élément photosensible est une photodiode. Comme illustré sur la Figure I.15, le nœud flottant du pixel est initialisé à une tension référence Vrst par le transistor M1. Puis, sous l'action du photocourant généré par la photodiode, les charges présentes sur ce noeud flottant vont être évacuées et le potentiel résultant sur la capacité va alors diminuer[12]. La pente de décroissance est proportionnelle à l'intensité lumineuse reçue par la photodiode. Il en est donc de même pour la différence de potentiel entre les deux instants d'échantillonnage[13]. Ainsi l'information utile est obtenue par double échantillonnage aux instants « t1 » et « t2 » puis différence, ce qui permet de s'affranchir du bruit spatial fixe dû aux dispersions des paramètres intrinsèques des composants d'un pixel à un autre[14] lors de la fabrication. A priori, on pourrait penser qu'il s'agit d'un double échantillonnage corrélé mais la nécessité de conserver un temps d'exposition oblige, pour ne pas avoir des cadences images trop lentes, à échantillonner d'abord la phototension, puis à procéder au reset (cf. Figure I.15).

Le modèle de photoréponse développé pour cette architecture de pixel par [Shcherback & Yadid-Pecht-04] est d'un grand intérêt car il permet d'optimiser le compromis existant entre le gain de conversion et la collection des photocharges. En effet, augmenter le gain de conversion signifie diminuer la capacité de la photodiode, c'est-à-dire en première approximation sa surface, et donc diminuer le nombre de photocharges collectées.

[12] Ce mode de fonctionnement est dit « à intégration du photocourant », sous entendu sur la capacité du nœud flottant, a été proposé initialement en 1967 par [Weckler-67].

[13] Au premier ordre, car en pratique la capacité de la jonction varie en fonction du potentiel à ses bornes [John&Martins-97].

[14] Notamment celles du transistor suiveur qui est la contribution majoritaire au bruit spatial fixe.

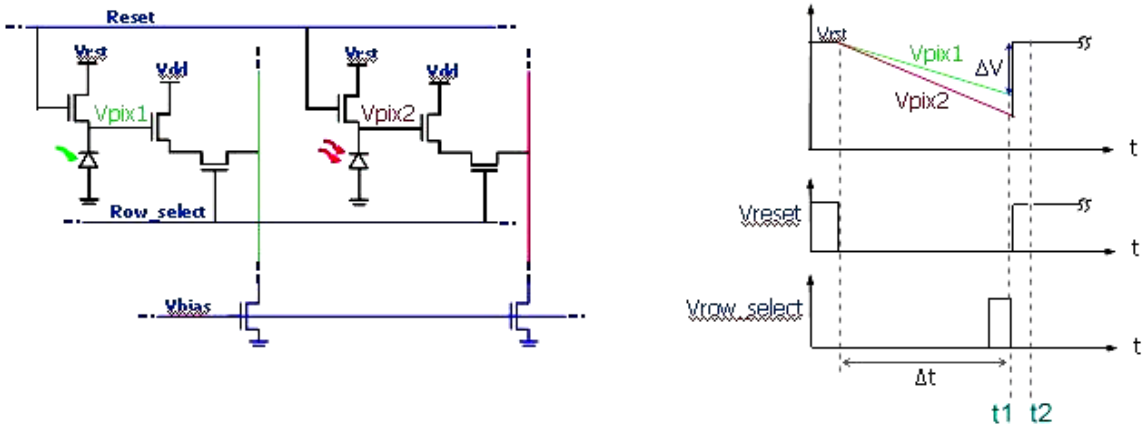


Figure I.15. Structure d'un pixel APS-3T (à gauche) et évolution du potentiel sur le nœud flottant de lecture au cours de la séquence pixel (à droite).

Le gain de conversion d'un pixel APS-3T se détermine en mode intégration du photocourant[15], il est de l'ordre de 2μV/e$^-$.

Cette architecture de pixel à trois transistors a constitué une avancée majeure en imagerie CMOS et a permis à ce type de capteurs de se développer et de prendre part au marché des capteurs d'images. Depuis quelques années, des structures de pixels actifs à quatre transistors sont exploitées et se généralisent car elles possèdent plusieurs avantages.

- APS-4T et APS-4TC

Le pixel actif à quatre transistors, schématisé Figure I.16, a été proposé par [Mendis et al.-93] afin d'obtenir une meilleure sensibilité et un bruit plus faible. L'APS-4T combine les performances photosensibles d'une photodiode verrouillée avec une réelle lecture à double échantillonnage corrélé (cf.Figure I.17) qui est la meilleure façon à ce jour de réduire le bruit pixel (cf. § III.2.b.) [Magnan et al.-04].

Le fonctionnement du pixel est le suivant (cf. Figure I.16 et Figure I.17) : le nœud flottant V_{ISF} étant fixé au potentiel VDD par le transistor de reset M_R, on décharge la photodiode à travers Tx puis on isole la photodiode pendant le temps d'intégration. On réalise une première lecture du pixel (avec bruit de reset), puis on transfère les charges accumulées dans la diode, par le transistor de transfert T_X, sur le nœud V_{ISF} pour lire ensuite le potentiel de ce nœud (signal utile avec bruit de reset et bruit lié au courant d'obscurité). La valeur finale du pixel est alors obtenue par différence de ces deux lectures, ce qui permet de réduire largement le bruit temporel de reset provenant de l'initialisation du pixel[16].

---

[15] *La capacité de conversion est dominée par celle de la photodiode, de l'ordre de 100 fF.*

[16] *La contribution du bruit en KTC est majoritaire en terme de bruit temporel dans les imageurs CMOS [Degerli-00]. Nous réalisons une synthèse des différents bruits à considérer dans les capteurs d'images lors du paragraphe Chapitre I.III.2.a.*

[Rhodes et al.-04] proposent d'insérer une capacité sur le nœud flottant de lecture (cf. Figure I.16) afin d'améliorer la linéarité et d'être capable par simple modification d'un masque, d'ajuster le facteur de conversion du pixel pour adapter et optimiser le capteur aux spécifications de sensibilité lumineuse requises[17].

Le **gain de conversion** de ce type de pixel APS-4T est de l'ordre de **30µV/e⁻**. Nous le déterminons par la capacité de conversion équivalente sur l'entrée de la grille du transistor suiveur après transfert des charges accumulées dans la photodiode verrouillée.

Cette structure est devenue le standard aujourd'hui en imagerie CMOS en raison du **faible courant d'obscurité** de son photodétecteur (**~40 pA/cm²**), d'un facteur de conversion élevé, mais aussi parce qu'il permet de réduire drastiquement la **surface des pixels** en partageant certains transistors entre pixels adjacents [Findlater et al.-03b] [Rhodes et al.-04].

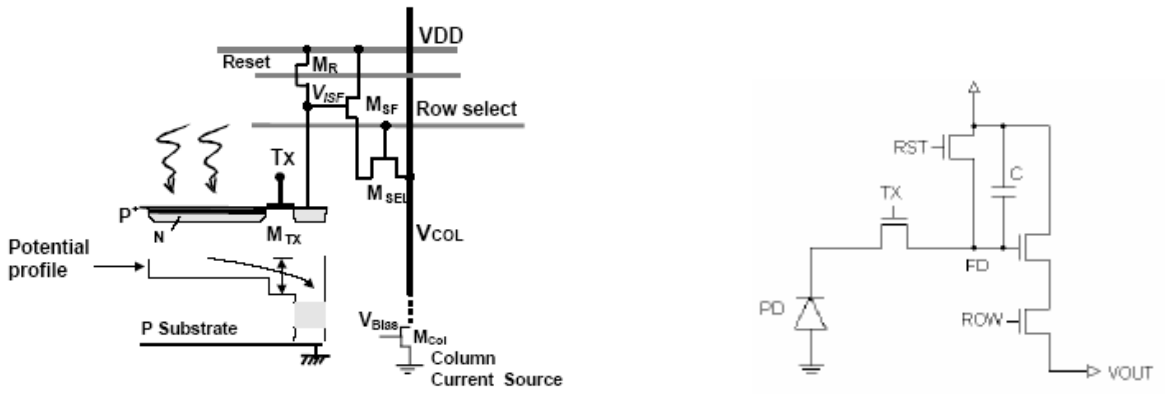


Figure I.16. Structure et profil de potentiel d'un pixel APS-4T [Magnan-03], et structure d'un pixel APS-4TC.

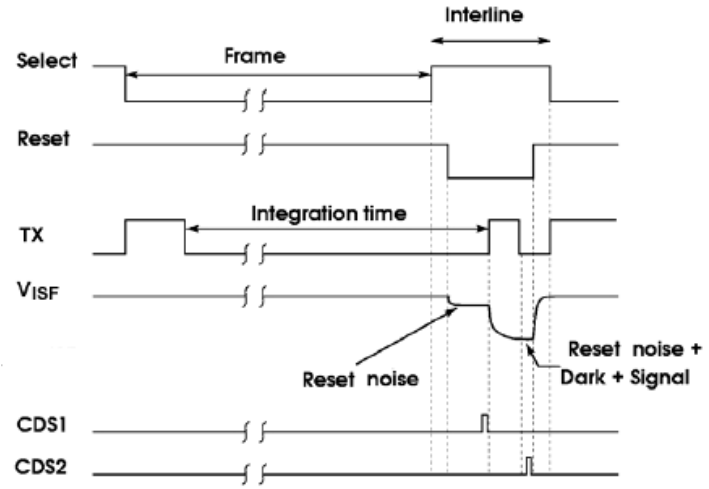


Figure I.17. Séquence de lecture d'un pixel APS-4T, avec un réel double échantillonnage corrélé.

[17] *Notamment le compromis entre gain de conversion important et faible flux lumineux, ou faible gain de conversion et capacité d'accumulation de charges.*

- APS-2.5T, APS-1.75T et APS-1.5T

La possibilité de **partage de transistors** de la structure APS-4T au sein de la matrice de pixel constitue un énorme atout puisque cela permet de réduire très efficacement leur taille (2 ×2μm² en procédé CMOS 150 nm [Kasano et al.-05]). Nous avons vu apparaître des pixels à 2.5 transistors par pixel, puis 1.75 transistors et même 1.5 transistors par pixel.

En effet, même si le transistor de transfert ne peut être partagé, les transistors de reset, de sélection et le suiveur sont mis en commun entre deux pixels consécutifs d'une même colonne pour obtenir un pixel à 1 + 3/2 = 2.5 transistors (cf. Figure I.18). Le mode de lecture est alors celui d'un pixel 4T à la différence près que la séquence de lecture, illustrée Figure I.17, est répétée deux fois consécutivement afin de lire les sorties des pixels des deux lignes partageant les trois transistors de lecture. La ligne supérieure est alors lue, puis c'est au tour de la ligne inférieure.

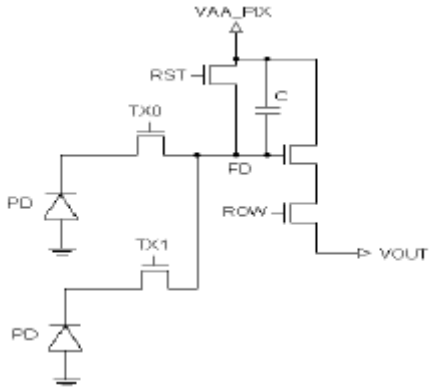


Figure I.18. Partage de trois transistors pour deux pixels d'une même colonne [Rhodes et al.-04].

Dans le cas de l'APS-1.75T, qui a été proposé par [Mori et al.-04], les transistors de reset, de sélection et le suiveur sont partagés par quatre pixels adjacents (cf. Figure I.19 gauche), ce qui équivaut à 1+3/4 = 1.75 transistors par pixel. La lecture s'effectue ici deux lignes par deux lignes en parallélisant la lecture sur deux bus signaux et en procédant séquentiellement au transfert des colonnes paires et impaires. Sur la Figure I.19 de droite, ce sont les deux lignes centrales qui sont lues : les charges accumulées dans les pixels « B » et « Gr » de la colonne impaire (« odd ») sont transférées en premier, puis vient le transfert des pixels « Gb » et « R » de la colonne paire (« even »).

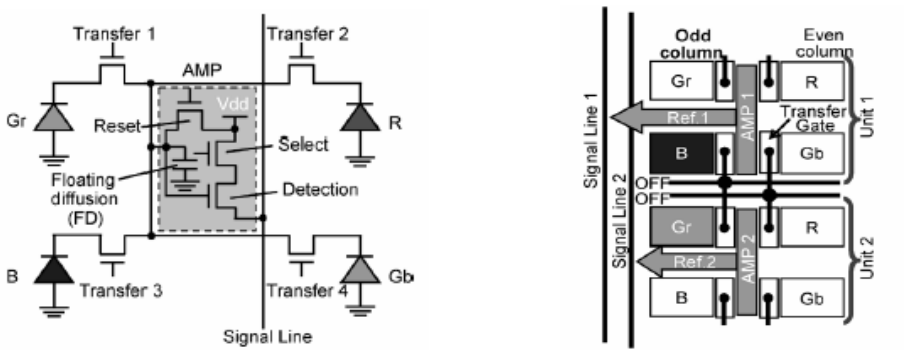


Figure I.19. Partage des trois transistors pour 4 pixels APS-1.75T (à gauche) et assemblage de deux groupes de pixels pour constituer la matrice entière (à droite) [Mori et al.-04].

Enfin, l'architecture de l'APS-1.5T [Kasano et al.-05] est représentée sur la Figure I.20. Le transistor de sélection de la Figure I.19 est supprimé et remplacé par une commande séquentielle du signal vdd. On obtient alors un pixel à 1+2/4 = 1.5 transistors. La contrainte de taille des pixels étant la priorité actuellement, cette architecture devrait se généraliser.

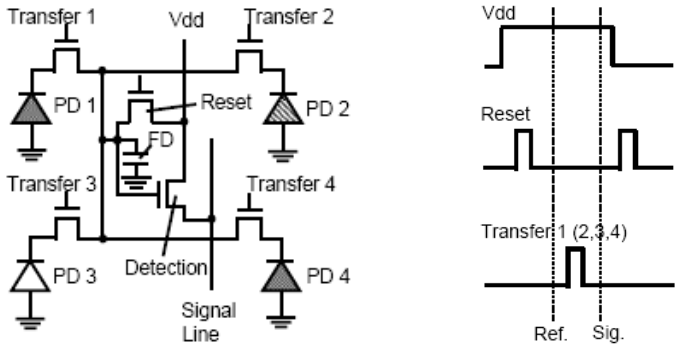


Figure I.20. Architecture et configuration de partage des transistors dans le cas de l'APS-1.5T, avec séquencement des signaux de commande [Kasano et al.-05].

III.1.b. Perspectives

La taille des pixels ne cesse de diminuer depuis toujours et l'on atteint aujourd'hui la taille critique de 2×2 μm² [Kasano et al.-05]. Le facteur de remplissage a lui aussi tendance à diminuer avec la finesse des technologies CMOS, comme l'illustre la Figure I.21, ci-dessous. On peut remarquer l'intérêt du partage des transistors, qui permet, à taille de pixel donnée, d'augmenter le facteur de remplissage, donc la capacité de collection du pixel et par suite les performances générales du capteur. Cependant ces performances sont globalement réduites avec la diminution de la surface du pixel [Theuwissen-02].

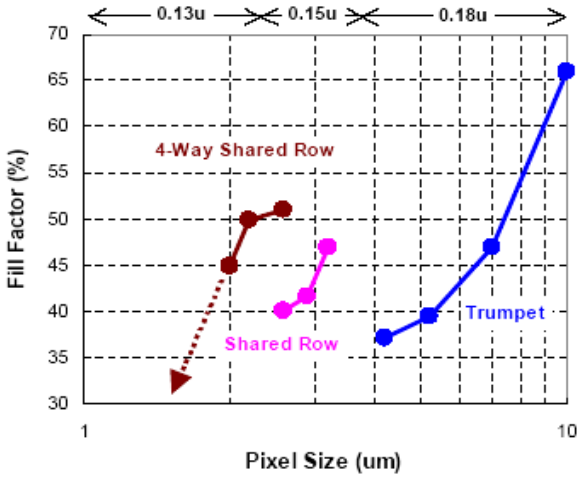


Figure I.21. Evolution du facteur de remplissage en fonction de la taille du pixel [Rodes et al.-04].

[Rodes et al.-04] montre aussi que les contraintes des technologies optoélectroniques au sein d'un pixel, aussi bien en technologie CMOS qu'en CCD, font qu'en l'état actuel des technologies, la taille des pixels ne pourra descendre au dessous de 2x2 µm² environ (cf. Figure I.22).

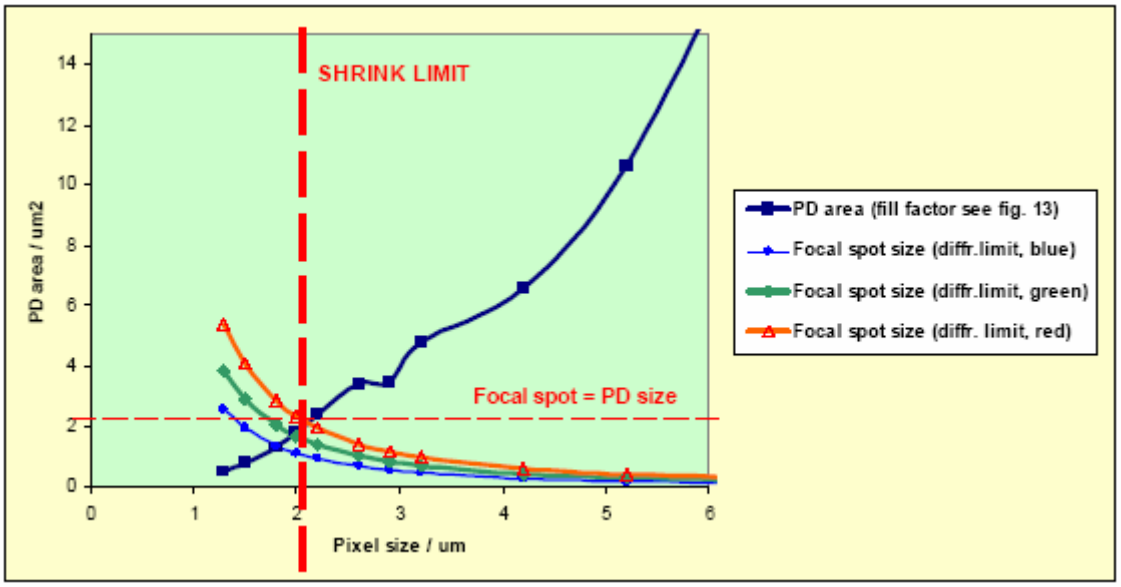


Figure I.22. Evolution des surfaces de photodiodes et des précisions de focalisations des micro-lentilles, et mise en évidence de la taille limite des pixels [Rodes et al.-04].

De plus, le procédé de fabrication CMOS des circuits VLSI « standards » possède environ une génération et demi d'avance sur le procédé de fabrication imagerie[18].

Ces deux aspects contribuent à l'émergence de nouvelles technologies d'intégration des capteurs d'images : en trois dimensions. Ces technologies permettent d'intégrer les photodétecteurs au dessus du procédé de fabrication CMOS (cf. Figure I.7 p. 13). Elles seront bientôt suffisamment maîtrisées pour accueillir les dispositifs d'imagerie de demain : photodétecteurs mais aussi photoémetteurs [Lule et al. -99a] [Theil-03]. Ces nouvelles applications de photodétection exploitent largement le domaine du visible, mais également de l'infrarouge [Dong et al.-05] ainsi que des rayons X pour l'imagerie médicale [Karim et al.-03].

Un autre avantage essentiel de dissocier le photodétecteur du circuit de traitement VLSI est de pouvoir bénéficier pleinement de l'augmentation de la précision lithographique CMOS sans être limité par les contraintes de surface dues à l'optique. Ainsi, de nouvelles architectures de conditionnement de l'information électrique photogénérée pourront être associées à chaque photodétecteur pour réaliser des

[18] *A la fin de l'année 2005 par exemple, STMicroelectronics va fondre ses premiers cœurs de processeurs en 90 nm alors que les imageurs en 130 nm ne seront vendus qu'en 2006.*

traitements dès l'acquisition lumineuse. [Lule et al.-00a] ont montré qu'il est possible d'intégrer, sur une surface pixel de taille 5×5 μm² et avec un procédé CMOS 100nm, 17 transistors et 2 capacités.

Avec de telles ressources de traitements, de nouvelles structures de pixels peuvent être envisagées. Des structures adaptatives, par exemple pour adresser des applications nécessitant une grande dynamique, devraient notamment autoriser des acquisitions de l'ordre de 120 dB [Benthien et al.-00]. Une spécification essentielle des capteurs d'images pour le marché automobile sera alors remplie.

III.2. Du pixel à l'image

Notre objectif final est de concevoir un imageur avec un rendu d'image et de vidéo optimum. Pour cela, il est important de considérer les différentes sources de bruits.

III.2.a. Bruit électronique dans un capteur d'image

Nous ne considérons pas ici les effets des différents composants optiques, et nous nous plaçons dans le cas d'une illumination idéale de l'imageur, sans distorsion optique. Deux types de bruits doivent alors être considérés : le bruit spatial fixe et le bruit temporel.

- Le bruit spatial fixe (« fixed pattern noise »).

Le bruit spatial est la différence de potentiel observée entre deux pixels distincts d'un même circuit soumis là aussi à une même illumination. Le bruit spatial fixe est lié aux dispersions des paramètres des composants apparaissant lors de la fabrication. Par exemple, d'un pixel à un autre, le transistor suiveur verra sa tension de seuil varier ce qui influera directement sur la tension de sortie du pixel.

- Le bruit temporel (« temporal noise »).

Le bruit temporel résulte de la différence de tensions entre deux mesures successives d'un même pixel sous illumination constante. Il provient quand à lui de la nature corpusculaire de la lumière et de l'électricité qui produisent des variations de tensions aléatoires, souvent modélisées par une loi de Poisson.

Les différentes contributions au bruit temporel proviennent par conséquent de chacun des éléments de la chaîne de traitement du signal (cf. Figure I.23).

Le photodétecteur est la première source de bruit par son courant photonique et d'obscurité. Les composants de l'électronique associée sont eux aussi caractérisés par diverses contributions en bruit qui sont : le bruit de grenaille (« shot noise »), en 1/f, et le bruit thermique (« thermal noise »). Ces différents bruits possèdent les caractéristiques suivantes :

- le **bruit de grenaille** (ou « shot noise ») résulte de la fluctuation du nombre de porteurs sous l'action d'un champ électrique. C'est une fluctuation qui suit la loi de Poisson. Ce bruit est donc directement lié au courant électrique auquel il se superpose. On le modélise par une source de courant, placée en parallèle avec le composant considéré, de densité spectrale de puissance :

Eq. I.17. $$Ig^2(f) = 2.q.I \quad (A^2/Hz)$$

où « q » est la charge élémentaire (1.6×10^{-19} C) et « I » est le courant moyen.

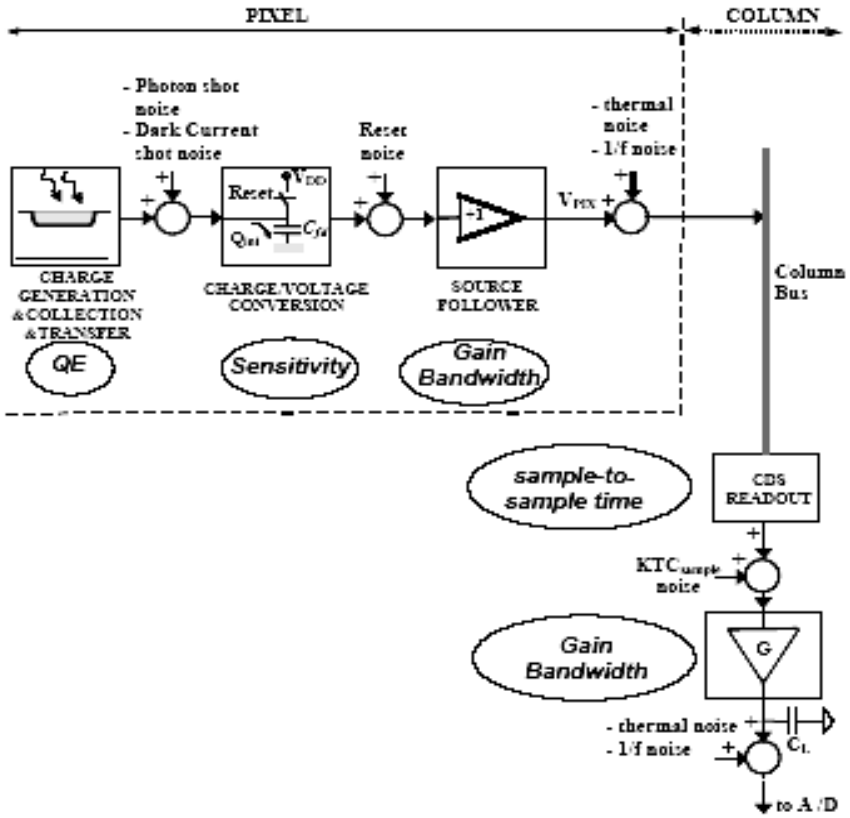


Figure I.23. Différentes contributions au bruit temporel dans un capteur d'images CMOS [Magnan-03].

- Le **bruit en 1/f** (appelé aussi bruit rose, ou scintillation, ou « flicker noise ») est un bruit dont la densité spectrale est inversement proportionnelle à la fréquence. Il est lié aux défauts du semi-conducteur (fluctuations des taux de recombinaison des paires électrons trous, piégeage des porteurs dans les défauts de l'oxyde du transistor), et dépend donc de paramètres technologiques. Dans le transistor MOS, il se manifeste comme une fluctuation de la tension de seuil. On le modélise alors par une source de tension placée en série avec la grille, de densité spectrale de puissance :

$$Vs^2(\mathrm{f}) = \frac{K}{W.L.C_{ox}.f} \quad (\mathrm{V^2/Hz}) \qquad \text{Eq. I.18.}$$

où « W » et « L » sont les dimensions du transistor, « K » est un coefficient empirique lié à la technologie, « Cox » est la capacité de l'oxyde mince et « f » est la fréquence.

Remarquons que, parmi les bruits que nous résumons ici, c'est le seul dont la densité spectrale dépend de la fréquence.

- Le **bruit thermique** (« thermal noise ») apparaît aux bornes de tout élément résistif. Il provient de l'agitation thermique des électrons qui engendre des fluctuations de la différence de potentiel aux bornes du composant. Sa modélisation la plus naturelle est une tension de bruit, de densité spectrale de puissance $V^2(\mathrm{f}) = 4.k.T.R \quad (\mathrm{V^2/Hz})$, placée en série avec la résistance. Où « k » est la constante de

Boltzman (1.38×10^{-23} J/K), « T » est la température (K), « R » est la valeur de la résistance et « h » la constante de Plank (6.626×10^{-34} J.s).

Ce générateur de Thévenin a un générateur de Norton équivalent dans lequel le bruit est modélisé par une source de courant, placée en parallèle, de densité spectrale de puissance $I^2(f) = \frac{4.k.T}{R}$ (A^2/Hz).

Dans une structure de type RC, la capacité ne génère pas de bruit, mais peut stocker le bruit thermique généré par la résistance. En considérant un système du premier ordre, la bande passante de bruit de ce filtre passe-bas de fréquence de coupure $f_0 = \frac{1}{2\pi.R.C}$ vaut $\frac{\pi.f_0}{2}$ [Johns & Martin-97]. Le bruit de la résistance est un bruit blanc, le bruit en sortie de ce système vaut alors :

Eq. I.19. $$V^2(f) = \frac{\pi}{2} \times \frac{1}{2\pi.R.C} \times 4.k.T.R = \frac{k.T}{C} \quad (V^2/Hz)$$

D'autres bruits sont liés à l'architecture du pixel, c'est le cas du **bruit de Reset** qui s'explique par deux phénomènes :

L'utilisation d'un transistor de type N comme transistor de reset dans les pixels actifs (APS-3T ou APS-4T), entraîne un offset négatif sur la tension mesurée aux bornes de la photodiode. Cet offset apparaît lors de la commutation du transistor de Reset (de '1' à '0'), il est dû aux injections de charges provenant du canal de ce transistor.

Lorsque le transistor de Reset est fermé, le système RC (transistor et capacité équivalente sur la grille du transistor suiveur) engendre un bruit $\frac{k.T}{C}$. Le transistor devient en effet sensible aux variations dues au bruit thermique à la fin de la précharge, car il est en mode de conduction sous le seuil. Ce bruit (aléatoire) entraîne une erreur lors de l'échantillonnage de la tension de précharge Vrst.

Il est reconnu qu'il faut minimiser non pas le bruit en sortie du capteur mais le bruit ramené en entrée (en électrons) qui est à comparer à la quantité d'électrons collectés. Les travaux de [Degerli-00] et [Magnan et al.-04] déterminent le bruit total ramené en charges en entrée d'un pixel APS-3T, il vaut :

Eq. I.20. $$Bruit\_total_{entrée} = \frac{\sqrt{\sigma^2_{V_O}}}{G_{conv}} = \frac{\sqrt{\sigma^2_{V_O}}}{A_{SF}.A_{COL}} \times \frac{C_{conv}}{q} \quad (e^-)$$

où « σ^2_{VO} » est la densité spectrale du bruit de la tension de sortie, « G_{conv} » est le gain de conversion, « A_{SF} » le gain du transistor suiveur du pixel, « A_{COL} » le gain de l'amplificateur de lecture colonne et « C_{conv} » la capacité de conversion.

Cette expression confirme l'importance du gain de conversion dans un pixel, mais aussi celles du transistor suiveur et de l'étage de lecture. En ce qui concerne le transistor suiveur, sa largeur de grille influe grandement sur la densité spectrale du bruit en sortie et on peut trouver une valeur optimale qui minimise le bruit total ramené en entrée.

III.2.b. Circuits de lecture du pixel

La conception du circuit de lecture est elle aussi critique car le niveau de bruit en dépend directement. La lecture ligne par ligne et en parallèle des pixels (cf.Figure I.25), avec un circuit de lecture par colonne, permet de réduire énormément la bande passante de ces circuits par rapport au cas d'un dispositif unique en sortie, comme c'est le cas des capteurs CCD. Par conséquent, leur contribution au bruit total est minimisée. De plus les contraintes de surface sont moins fortes en bout de colonne qu'au niveau pixel ce qui autorise une meilleure optimisation lors de la conception.

Un circuit réalisant l'opération de double lecture est représenté Figure I.24. Dans le cas de l'APS-3T, ces deux instants ne peuvent appartenir à la même pente d'intégration (cf.Figure I.15 page 24), le bruit de reset subsiste et seul le bruit spatial fixe du pixel et celui en 1/f du transistor suiveur peuvent être éliminés.

Par contre, les pixels actifs à quatre transistors APS-4T, qui séparent le photodétecteur du lieu de conversion charge-tension, autorisent ce double échantillonnage tout en conservant le même bruit de reset : l'opération est alors appelée à double échantillonnage corrélé. De cette manière, la différence des deux potentiels échantillonnés fournit le signal utile et le bruit lié au courrant d'obscurité. Le rapport signal sur bruit se trouve ainsi augmenté de 40 dB environ.

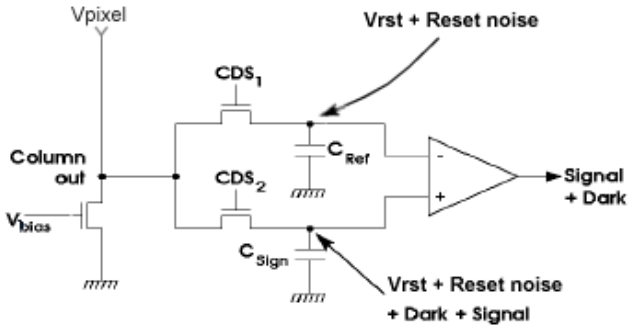


Figure I.24. Circuit de lecture réalisant le double échantillonnage d'un pixel.

Le double échantillonnage corrélé élimine une grande part du bruit issu de l'électronique de conditionnement des charges photogénérées. Cependant il ne supprime pas le courant d'obscurité (cf. Figure I.24). De plus, si un photodétecteur ne possède pas le même rendement de détection que son voisin, cela ne sera pas compensé et produira une non uniformité spatiale du capteur. Pour atténuer ces défauts, une approche courante consiste à réaliser une acquisition d'image « noire », sans flux lumineux incident, et de soustraire cette image à celles normalement acquises. Ceci est réalisé par post-traitement au moyen d'un coprocesseur. Dans la section suivante, nous décrivons l'architecture du système sur puce global assurant la capture et la mise en forme des images.

III.2.c. Architecture système d'un imageur, création de l'image

A ce niveau de la conception du capteur d'images, nous sommes capables d'échantillonner la scène à un instant donné (pixel) et d'en extraire les différences de potentiels constituant l'information sur la quantité

de lumière reçue (circuits de lecture). Il reste désormais à réaliser cette lecture pour toute la matrice de pixels, et à gérer ensuite l'ensemble de ces données pour restituer une image ou un flux vidéo.

L'acquisition d'une image complète est obtenue par lectures successives des lignes de la matrice de pixels, il s'agit d'un mode de lecture « progressif ». Les lignes sont adressées une à une par un décodeur de lignes (cf. Figure I.25) pilotant le transistor de sélection des pixels. Le temps d'exposition des pixels à la lumière incidente, commun à tous les pixels, est obtenu par des signaux d'initialisation de lignes que l'on distribue sur les grilles des transistors de reset. Une ligne étant sélectionnée, les tensions de sortie des pixels sont lues parallèlement par les circuits de lecture puis numérisées et mémorisées dans un plan mémoire de type SRAM.

Nous disposons alors d'une image brute qui contient du bruit (le courant d'obscurité et les non uniformités spatiales) et dont les caractéristiques visuelles peuvent être sensiblement améliorées par un post-traitement. Celui-ci consiste à réduire le bruit et à corriger les pixels défectueux, à améliorer les contrastes, mais aussi à corriger les anomalies lumineuses introduites par le conditionnement optique (distorsions, ombres, différences de focus sur la périphérie).

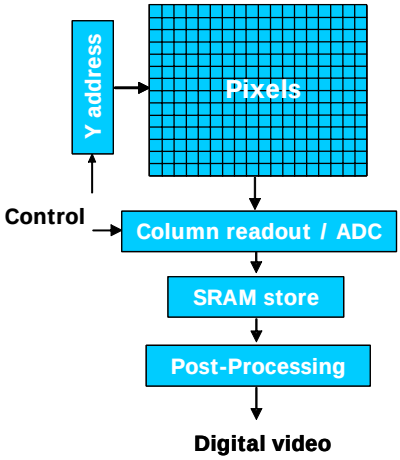


Figure I.25. Architecture système d'un imageur.

De plus, la majorité des imageurs restituent aussi la couleur. Dans ce cas, chaque luminance pixel correspond à une bande chromatique du motif de Bayer qu'il faut interpoler avec les deux autres bandes chromatiques contenues dans les pixels voisins pour reconstituer l'information de couleur complète (Rouge-Vert-Bleu). Une image couleur nécessite a priori trois fois plus de mémoire que son équivalent monochromatique. Cependant, la perception de la couleur étant moins importante que celle de la luminance en vision humaine, un codage dissociant ces deux informations est souvent préféré pour favoriser l'information de luminance par rapport à la couleur qui est alors sous échantillonnée[19]. D'autres opérations, comme la balance des blancs, peuvent elles aussi être menées lors de ce post-traitement.

[19] *De plus lorsqu'il est appliqué avant la compression JPEG, ce codage de l'information pixel permet d'obtenir un meilleur taux de compression.*

Enfin, pour réduire la consommation du système et la quantité de données à enregistrer ou transmettre, il est essentiel de compresser les données pixels. Cette étape est elle aussi intégrée dans ce post-traitement.

Cet ensemble de tâches est effectuée à l'aide d'une architecture digitale dédiée : le coprocesseur d'images. Un exemple d'une telle architecture de système sur puce d'imagerie complet est donné sous forme de schéma blocs en Annexe C.

III.2.d. Performances typiques d'un imageur

Nous résumons dans cette section les caractéristiques électro-optiques qui définissent un imageur, avec leurs valeurs typiques.

CARACTERISTIQUES	VALEURS TYPIQUES
Résolution	1.2 megapixels (1280×1024)
Taille pixel	3×3 μm²
Facteur de remplissage	35%
Sensibilité[20]	5000 e^-/lux.s
Gain de conversion	35 μV/e^-
Courant d'obscurité	40 pA/cm²
Capacité de charges à saturation[21]	8000 e^-
Bruit temporel ramené en entrée	8 e^-
Dynamique[22]	60 dB
Bruit spatial fixe, non-uniformité	100 ppm
Puissance consommée (analog et digital)	100 mW
Temps d'exposition à cadence maxi	2μs à 30ms @ 30 im/s
Précision / Format de pixel	11bits
Cadence image	1 à 30 im/s

Tableau I.2. Caractéristiques électro-optiques d'un imageur.

III.3. L'imageur, un système sur puce dans un contexte concurrentiel

Le seul secteur de la téléphonie a représenté pour les capteurs d'images un marché de 3.4 milliards de dollars en 2004 [Future Horizons-05]. Il s'agit donc d'un marché particulièrement conséquent et donc très concurrentiel où la contrainte de coût est forte. Pour être compétitif, le prix d'un capteur embarqué dans les appareils portables comme les assistants personnels ou les téléphones mobiles doit avoisiner les 10 euros.

[20] *La sensibilité caractérise la qualité de l'ensemble électro-optique en informant sur la quantité de charges effectivement générées par le photodétecteur pour un flux lumineux donné.*

[21] *Nombre total de photocharges pouvant être accumulées dans le photodétecteur.*

[22] *Rapport entre la quantité de charges utiles que peut traiter le capteur et la quantité de charges de bruit qui le caractérise.*

La réduction du prix de revient se traduit au niveau intégration silicium par la minimisation de la surface requise. En effet, pour un imageur de type mégapixels, qui devient la résolution standard aujourd'hui, la part associée au silicium constitue plus du tiers du coût total (cf. Figure I.26).

Ainsi, minimiser cette surface est un impératif et ce même si la part de l'encapsulation et de l'ajustement optique du système intégré tient une place de plus en plus importante dans cette répartition du coût de fabrication.

Nous mettons en évidence cette tendance sur la Figure I.26 où l'on remarque que l'évolution des lithographies CMOS permet de réduire le coût relatif au silicium requis pour intégrer des systèmes de plus en plus complexes. C'est l'aspect intégration de systèmes hétérogènes, appelé aussi « System in Package », qui devient prédominant en termes de coût de fabrication, mais aussi d'un point de vue technique où de nouveaux challenges technologiques apparaissent[23].

Actuellement ce sont l'optique et le post-traitement des images acquises qui sont critiques pour la qualité des images restituées par les modules imageurs des téléphones mobiles (Figure I.27). Ainsi il est essentiel, pour un concepteur et fabricant d'imageurs tel que STMicroelectronics, d'avoir la maîtrise de la chaîne vidéo complète pour être capable d'optimiser les performances globales et le coût. Cela comprend des savoirs faire en conception électronique, en dépôt de résine (microlentilles), en optique générale, ainsi que pour l'encapsulation du module complet.

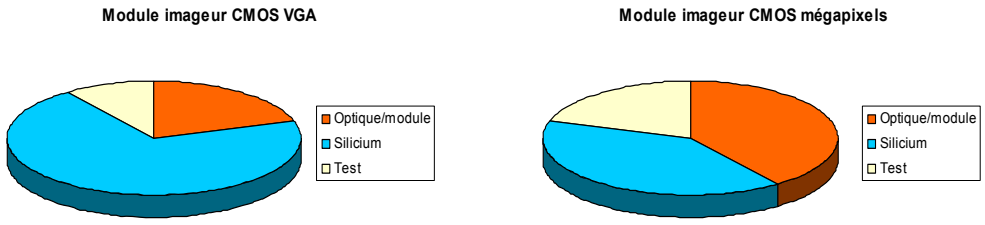


Figure I.26. Répartition du coût de fabrication d'un module imageur CMOS pour une résolution VGA (lithographie 350 nm) et Megapixels (lithographie 180 nm).

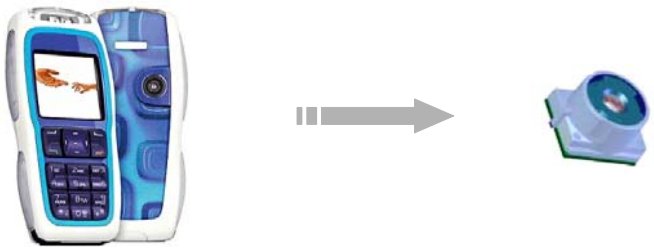


Figure I.27. Module imageur embarqué dans un téléphone mobile.

[23] *L'empilement de plusieurs puces dans un seul boîtier par exemple (Philips).*

Cette évolution vers des systèmes « tout intégré » n'a pas pour raison l'amour de l'art, mais bel et bien les contraintes de coût. D'un point de vue VLSI, Texas Instruments propose par exemple un circuit pour le standard 2.5G embarquant l'essentiel du traitement du signal d'un combiné GSM/GPRS. Cela signifie que l'émission et la réception RF, le traitement bande de base analogique et numérique, et la mémoire SRAM sont « on-chip ». Seule une part de la fonction d'alimentation, l'amplificateur de puissance et la mémoire flash requièrent des composants externes. La technologie de fabrication est une CMOS 90 nm, pour une puce nécessitant moitié moins d'énergie et d'encombrement qu'une approche traditionnelle à deux ou trois circuits [Elec. Int.-05].

Cependant, il existe souvent un décalage entre l'offre d'un système sur puce tout intégré et la demande de nouvelles fonctionnalités. Une étape intermédiaire est alors ajoutée en passant par un circuit dédié. A l'heure de la vidéo dans les appareils mobiles, il n'existe pas encore de standard reliant le processeur bande de base à un coprocesseur d'applications ou à un imageur avec traitement du signal intégré. STMicroelectronics et Texas Instruments proposent OMAPI, alors que d'autres sociétés vantent d'autres standards.

Ces préoccupations d'intégration d'une fonctionnalité en minimisant le coût tout en optimisant les performances constitueront les critères déterminants de notre choix d'architecture.

IV. CAPTEURS POUR LA VISION ARTIFICIELLE, LES RETINES

Les capteurs que nous avons décrits jusqu'ici, les imageurs, sont par définition optimisés pour l'acquisition d'images à des fins d'imagerie. Ils sont le fruit de plusieurs décennies de recherches[24] avec pour objectif de restituer la meilleure image possible.

Cependant l'acquisition d'images ne correspond pas forcément au besoin applicatif et l'analyse visuelle de la scène est parfois nécessaire, comme c'est le cas pour la reconnaissance d'objets, la mesure de distance, la détection de mouvement par exemple. Ces dernières peuvent être menées en procédant à un traitement des données numériques des pixels provenant de l'imageur. Pour cela une architecture de traitement s'appuyant sur un co-processeur, un DSP, ou un FPGA pourra être choisie. L'architecture embarque alors un algorithme développé et préalablement validé.

Dans la plupart des cas, cette approche de traitement du signal permet d'aboutir à la fonction souhaitée, mais parfois au prix d'une charge de calculs considérable. De fait, lorsqu'il s'agit d'embarquer l'application dans des dispositifs portables où la quantité d'énergie et les ressources matérielles sont sévèrement limitées, l'intégration de la fonctionnalité n'est plus possible.

Il faut cependant garder à l'esprit que, lorsque l'on opère un échantillonnage spatial et temporel de la scène observée, il s'ensuit nécessairement une perte d'information. Or, cette perte peut contenir des données essentielles pour accomplir la tâche de vision souhaitée. Ceci va engendrer la consommation

[24] *Les premiers travaux sur les capteurs d'images datent des années 60 [Fossum-97].*

d'une certaine quantité d'énergie et de ressources matérielles pour reconstituer ou tenter de reconstituer l'information perdue a priori.

La manière de percevoir l'information lumineuse, donc l'acquisition et le conditionnement de l'information photoélectrique, constitue une étape cruciale dans le processus de traitement du signal à mettre en œuvre. Cependant, une modification de cette étape remet en cause une grande partie du traitement du signal qui suit, donc de l'architecture associée, ce qui est très coûteux en temps de conception et de validation. Ajouté aux progrès en densité d'intégration CMOS et à la réduction du temps de mise sur le marché, « time to market », cela mène les industries à adopter une architecture « imageur + traitement » pour la très grande majorité de leurs produits.

Pourtant, la communauté scientifique, à laquelle nous faisons partie, s'intéresse à des architectures de capteurs d'images pour lesquels la manière de conditionner le signal photoélectrique perçu fait partie intégrante du traitement du signal à mettre en œuvre pour accomplir la tâche de vision souhaitée.

Les rétines électroniques sont par définition[25] de tels dispositifs qui sont optimisés dans leur ensemble pour répondre à une tâche donnée. Il s'agit de l'approche « capteur intelligent », encore appelée « smart sensor », telle que l'a introduite Peter Burt en 1988 [Burt-88] [Jolion-01].

Les traitements au niveau pixel exploitent le caractère massivement parallèle mis en place pour réaliser l'échantillonnage spatial du flux lumineux. Le temps de traitement est alors indépendant du nombre de pixels, ce qui n'est pas possible dans le cas d'architectures numériques de post-traitement des données pixels.

IV.1. Conditionnement spécifique de l'information électrique, le pixel « intelligent »

L'architecture des pixels d'une rétine peut se caractériser par un fonctionnement synchrone ou asynchrone, une tension de sortie continue ou discrète, et une interdépendance ou non avec les pixels voisins. Ces architectures sont toutes destinées à un traitement dit « bas niveau » puisqu'il s'opère directement sur la luminance des pixels.

Selon les spécifications de l'application, telle ou telle structure de pixel pourra présenter un intérêt particulier. Nous décrivons dans la suite de ce paragraphe un état de l'art sur les différentes architectures de pixels existantes. Pour cela, on distingue celles dédiées à des traitements spatiaux de celles dédiées à des traitements temporels. Le cas spatio-temporel sera spécifiquement traité au chapitre suivant. Le champ applicatif étant large, les critères de performances pour de tels pixels ne peuvent pas être généralisés. On rencontre alors dans la littérature des critères spécifiques à chaque besoin, plus ou moins précis et objectifs, que nous ne décrirons pas ici.

[25] *« Analyser l'image là où elle est acquise pour n'en retenir et en transmettre qu'un extrait pertinent pour la tâche de vision en cours, tel est le principe des rétines artificielles. » est la définition donnée par [Bernard-97].*

IV.1.a. Traitement spatial

Dans le cas d'un traitement spatial du signal électrique photogénéré, plusieurs pixels de la matrice sont mis en jeu pour réaliser, par exemples, des opérations de filtrage, de transformées d'images ou d'égalisation d'histogrammes.

Ces traitements spatiaux s'obtiennent par opérations élémentaires comme la somme et la différence, l'agrégation[26], la valeur absolue, la comparaison, le maximum et le minimum, le produit, ou le logarithme et l'exponentielle des luminances des pixels. Du point de vue de leur conception, ces opérations sont obtenues à partir de briques de base dont l'information provenant du photodétecteur est traitée soit en courant, soit par transfert de charges, soit en tension.

Dans le premier cas, les photocourants sont conditionnés à l'aide de structures telles que les miroirs de courant, les convoyeurs de courant, les réseaux résistif[27], ou d'autres structures qui sont décrites dans [Mead-89] et [Vittoz-94] notamment. Il s'agit ici d'un fonctionnement en temps continu, avec des données continues elles aussi. Nous donnons dans le tableau, ci-dessous, l'encombrement à prévoir, en termes de transistors et de surface rapportée à la précision lithographique λ, pour chaque opération élémentaire.

Opération	*Nb transistors*	*Surface, f(λ)*	*Surface, λ=0.13µm*
Somme / différence	*2 / 4*	*12 λ^2 / 24 λ^2*	*0.2 µm² / 0.4 µm²*
Agrégation 1d / 2d	*1 / 2*	*6 λ^2 / 12 λ^2*	*0.1 µm² / 0.2 µm²*
Valeur absolue	*3*	*18 λ^2*	*0.3 µm²*
Maximum / Minimum	*5 / 7*	*30 λ^2 / 42 λ^2*	*0.5 µm² / 0.7 µm²*
Log / Exp	*1*	*6 λ^2*	*0.1 µm²*
Produit	*9*	*54 λ^2*	*0.9 µm²*

Tableau I.3. Encombrement lié à un traitement analogique au niveau pixel (en courant), (architectures décrites dans [Mead-89] et [Vittoz-94]).

Dans le cas de traitements en transfert de charges, les structures à mettre en œuvre fonctionnent en temps discret et sont à capacités commutées, comme présenté dans [Umminger & Sodini-92] [Ni et al.-93].

---

[26] L'agrégation de signaux est une qualité importante de notre système nerveux qui permet de signaler un évènement uniquement si un ensemble de signaux élementaires répondent à un critère donné. Cette caractéristique est souvent mise à profit au sein des rétines afin d'optimiser le temps et les ressources de calculs disponibles en ne considérant que les données importantes parmi l'ensemble. Une somme de courants constitue par exemple une aggrégation [Mead-89].

[27] L'intégration silicium des résistances étant réalisée par des transistors MOS en régime linéaire et également en faible inversion afin de limiter la consommation.

Les architectures de traitement en tension sont plus communes et peuvent être des amplificateurs, des comparateurs, ou des multiplieurs [Ruedi et al.-03].

On peut également noter que certaines technologies, comme les MOS à double grille utilisés en technologie FLASH, sont intéressantes pour concevoir des structures à gains programmables[28]. Les réseaux résistifs (ou capacitifs) permettent alors d'effectuer des opérations de filtrage spatial, comme le décrit [Bandyopadhyay et al.-06] pour une application de codage JPEG et motion-JPEG.

Nous présentons, ci-dessous, un exemple démonstrateur de l'intérêt de tels traitements de bas niveau. Il s'agit d'une simple opération de différence de luminances entre photodétecteurs voisins, permettant d'extraire les formes de la scène observée. En effet, comme l'illustre la Figure I.28 ci-dessous, nous sommes capables de reconnaître le contenu de la scène uniquement en détectant les contrastes[29] de l'image et en n'affichant que les points de l'image qui possèdent une différence de luminance supérieure à 15% de la dynamique de codage de l'image (8bits par exemple). Ces contrastes résultent de la différence entre la luminance du photodétecteur courant et celle de son voisin supérieur (masque de convolution $M = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$).

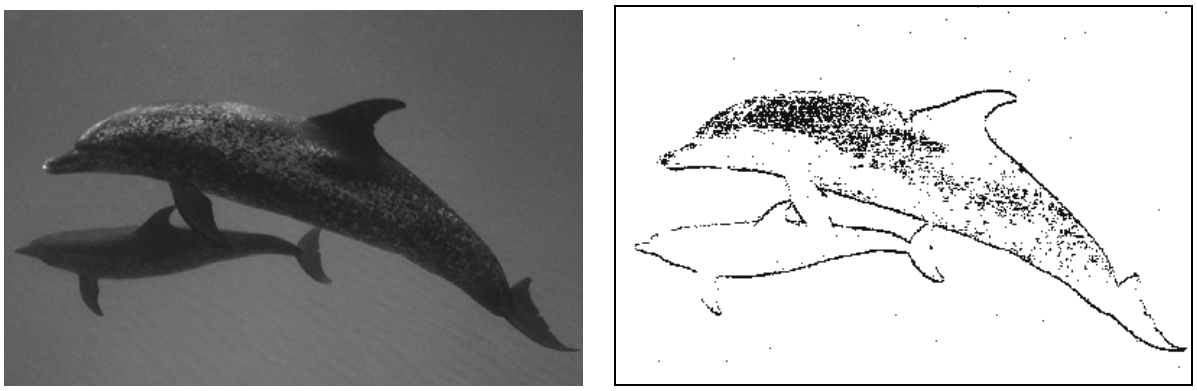


Figure I.28. Détection de contrastes horizontaux par différences de luminances verticales. Seuls les contrastes supérieurs à 15% de la dynamique de codage ne sont reportés dans le cadre de droite, soit 2,3% du nombre de pixels.

[28] *Ils sont également utilisés pour réaliser des traitements analogiques de précision. Ils peuvent permettre par exemple de compenser les tensions de décalages ou les défauts d'appariements des amplificateurs. En effet, on ajuste dans ce cas la tension de seuil des transistors de la paire différentielle de manière à obtenir une sortie nulle pour une tension différentielle d'entrée nulle aussi.*

[29] *Le contraste entre deux pixels peut se définir par le rapport :* $C = \frac{Pix1 - Pix2}{Pix1 + Pix2}$

L'ensemble des points affichés ne représente pourtant que 2,3% des pixels de l'image de départ, et ont été obtenus par un traitement d'image de bas niveau et de très faible complexité, dont l'intégration silicium requiert typiquement 5 transistors par pixels[30].

Les traitements spatiaux s'avèrent utiles dans de nombreux cadres applicatifs dont nous donnons quelques exemples dans la section IV.2.a.

IV.1.b. Traitement temporel

Les opérations au niveau pixel peuvent également renseigner sur le contenu temporel de la scène. Dans ce cas le fonctionnement de chaque pixel est indépendant de ses voisins. Comme dans le cas précédent d'un traitement spatial, on peut opérer soit en temps continu, soit en temps discret.

En continu, on ne procède pas nécessairement à une période d'intégration du photosignal pour procéder au traitement. En accord avec notre description du bruit dans les imageurs, le bruit photonique ne sera alors pas filtré et son influence sera à considérer avec attention lors de la conception et de la validation des pixels. Les structures à mettre en place ici sont alors généralement des filtres analogiques, actifs ou passifs, permettant de favoriser des variations lumineuses lentes ou rapides.

Des traitements temporels peuvent aussi être menés en temps discrets, en utilisant une période d'intégration pour réduire le bruit photonique et conserver un bon rapport signal sur bruit.

IV.2. Quelques applications

Depuis plus d'une quinzaine d'années, diverses équipes de recherche et laboratoires ont développé des rétines [Moini-99], bien souvent dans le cadre de recherches. Il est parfois moins connu que certains de ces circuits photosensibles ont remporté un franc succès dans le secteur industriel, pour des applications grand public par exemple. Nous faisons état dans cette section de quelques uns de ces prototypes ou produits développés qui renseigneront à la fois sur les applications possibles ainsi que sur les caractéristiques et performances générales obtenues[31].

Il est important de mentionner qu'un traitement temporel peut être associé à un traitement spatial pour renseigner sur l'évolution spatio-temporelle des constituants d'une scène. Il s'agit alors de la perception du mouvement dans la scène observée. Nous consacrons le chapitre suivant exclusivement à cette application.

---

[30] Cette structure à 5 transistors est mise en œuvre par [Nishio et al.-06] pour effectuer une détection de contrastes en réalisant une différence de courants au sein du pixel.

[31] La taille pixel, caractéristique importante, sera notamment exprimée relativement à la précision lithographique.

IV.2.a. Applications avec traitement spatial

Des rétines intégrant un traitement spatial ont été développées pour des applications nombreuses et variées. Nous évoquons ici la perception de scènes en 3D, la mesure de focus ou de mise au point optique, la perception du centre de gravité 2D, la reconnaissance d'objets et de formes, l'imagerie avec extraction de contraste, le filtrage, la détection du mouvement et le suivi d'objets.

Deux familles de techniques existent pour percevoir la scène en 3D : la stéréovision ou l'acquisition d'image avec lumière structurée, et la mesure du temps de parcours d'une émission laser sur la scène à percevoir. La première technique, la stéréovision, a été mise au point par une équipe de R&D de Mitsubishi Electric [Kage et al.-98]. Pour cela, deux rétines de résolution 32×32 et 256×256 pixels sont utilisées [Kyuma et al.-97]. Elles fonctionnent en mode intégration du photocourant, avec sortie en courant, capables d'une vitesse d'acquisition vidéo jusqu'à 1000 im/s et jouissent de fonctions de fenêtrage, d'extraction de contours, de filtrage spatial (moyenneur), et d'estimation du flot optique. A partir des flots optiques perçus par les deux rétines, les auteurs mettent en correspondance ces informations de mouvements pour en déduire le mouvement 3D des objets. L'application à la reconnaissance des mouvements de tête d'un individu est décrite avec succès, mais il s'agit d'une perception qualitative[32].

La mise au point d'un dispositif optique s'obtient communément en analysant les saillances des images. En effet, du point de vue de l'analyse spectrale de l'image reçue, l'optique est correctement mise au point lorsque l'on se place sur le maximum d'énergie contenu dans les hautes fréquences spatiales de l'image. [Delbruck-00] propose une rétine capable d'extraire en temps continu les contrastes et d'en faire la somme pour finalement restituer un courant qui est fonction de la quantité totale de contraste perçu. La mise au point est alors obtenue au maximum du photocourant. La taille des pixels est de 50x50 λ² (soit 6.5 x 6.5 µm² en process CMOS 0.13um), pour une matrice de 25x26 pixels, et une consommation de 0.5 mW.

[Clapp & Etienne-Cummings-02] ont quant à eux développé une rétine à pixels hybrides, capables à la fois d'acquérir des images à l'aide de pixels de type APS-3T et de restituer le barycentre de pixels détectant une variation lumineuse (objets en mouvement). Ces pixels spécifiques possèdent une conversion lumière-tension logarithmique et intègrent un comparateur qui impose un état haut en sortie lors d'une variation de lumière. La taille du pixel centre de masse est de 58.8x58.8 λ² (soit 7.8x7.8 µm² en CMOS 0.13 µm), pour une consommation de 0.38 mW. La vitesse de restitution des coordonnées du centre de masse dans le plan image est comprise entre 180 et 3580 points par seconde.

La reconnaissance d'objets et de formes est elle aussi possible par l'approche rétine. [Cathebras et al.-02], ont conçu une rétine de 100x100 pixels calculant en temps continu et en valeur continue (courant de sortie) la fonction d'intercorrélation entre l'image courante et une image de référence apprise et mémorisée

[32] *A la différence de la perception trois dimensions que nous décrivons dans la section suivante.*

au début de l'expérience. Les pixels fonctionnent en mode courant et sont de taille 50.6×50.6 µm² en technologie 0.6 µm, c'est-à-dire 84x84 λ² (soit 10.9x10.9 µm² en CMOS 0.13µm).

Des opérations de filtrage des images, des filtres de type median/moyenneur, exponentiel, gaussien, ou laplacien peuvent être mis en œuvre pour réduire le bruit sur l'image ou extraire la texture de la scène [Ni et al.-96]. Les formes, les contours et les bords des objets constituent par exemple les informations de sortie du circuit d'extraction de [Andreou & Boahen-95]. Le circuit comprend une matrice de 230×210 pixels, qui fonctionnent essentiellement en régime d'inversion faible et imitent les propriétés biologiques de perception des contrastes et d'adaptation aux conditions lumineuses. Le circuit occupe une surface de 9.5×9.3 mm² en technologie 1.2 µm, et chaque pixel est de taille 66×73 λ².

[Ruedi et al.-03] ont développé récemment un capteur d'extraction des contrastes fonctionnant sur une dynamique de lumière ambiante de 120 dB et transmettant les pixels contrastés de façon asynchrone et ordonnée dans le temps. Le pixel d'amplitude la plus élevée est transmis en premier, et celui de plus faible amplitude en dernier. La taille des pixels est de 69×69 µm² en technologie 0.5 µm, soit une taille relative de 138×138 λ².

IV.2.b. Applications avec traitement temporel

L'analyse temporelle intra-pixels du signal lumineux a servi plusieurs applications telles que la perception 3D, l'analyse temporelle de la lumière d'une scène, ou la compression vidéo.

Si ce traitement temporel peut être réalisé en temps continu et intégré dans chaque pixel comme c'est le cas des pixels adaptatifs de [Delbruck-04] [Kramer-02] [Delbruck & Mead-96], il peut aussi être obtenu en temps discret [Pain-01] [Gruev & Etienne-Cummings-04].

Comme nous l'avons évoqué, la perception 3D de la scène peut être obtenue par mesure de l'intervalle de temps entre l'émission d'une lumière laser et celui de sa réception sur la rétine après réflexion sur la scène observée. [Viarani et al.-04] mettent en œuvre une illumination continue et uniforme de la scène par trains d'impulsions à l'aide d'une LED de longueur d'onde 880 nm. La précision atteinte est de 15cm sur des distances de 3 à 9 m. La taille du pixel est de 180×160 µm² (soit 300×266 λ²) incluant une photodiode N-well / P-sub de taille 64×64 µm².

Certaines scènes comportent des objets ou sources lumineuses modulées dont les variations d'intensités sont périodiques (un ventilateur en rotation par exemple). [Sutherland et al.-02] proposent de caractériser le spectre de ces sources en intégrant dans chaque pixel logarithmique un filtre de type passe-haut puis un comparateur afin de générer un signal périodique à la fréquence du signal lumineux. Les mesures menées à partir d'une source de type LED infrarouge et sous lumière intérieure néon ont montré que le meilleur gain du phototransistor par rapport aux photodiodes (+5dB ici) était bénéfique pour l'application puisque l'indépendance des pixels fait que les dispersions de gain d'un phototransistor à un autre ne sont pas préjudiciables. En revanche, un comparateur à hystérésis initialement inséré, pour augmenter la robustesse de la comparaison, a finalement dégradé les performances à cause des faibles amplitudes de tensions d'entrée. Celles-ci rendaient alors le comparateur insensible aux variations.

Un intérêt applicatif important d'un traitement temporel dans le pixel peut être la compression des informations vidéo, en encodant uniquement les variations de luminance dans la scène, c'est-à-dire les contrastes. Nous avons relevé la conception de [Ruedi et al.-03]. Une démonstration de l'intérêt d'une transmission de ce type est visible sur le site web de [Delbruck] où un imageur extrayant les contrastes en temps continu est présenté [Delbruck & Mead-91]. Les pixels de cet imageur sont bio-inspirés [Delbruck-89] et ne sont sensibles qu'à des variations rapides du flux lumineux reçu. En effet, comme nous le décrivons dans la section II.1.b. du chapitre II, les rétines biologiques sont constituées de cellules dites bipolaires qui se comportent ainsi. Ces pixels sont particulièrement intéressants car, outre leur comportement dérivateur temporel, ils sont auto-adaptatifs. C'est-à-dire que chaque pixel s'adapte à la luminosité ambiante de manière à conserver son caractère dérivateur.

La compression vidéo ainsi que certaines applications telles que la segmentation par détection de contours, la détection de mouvement ou encore son estimation, mettent bien souvent en jeu la différence de luminance entre deux images consécutives. Afin de réduire la charge de calcul globale, [Pain et al.-01] ont mis au point une structure de pixel permettant de restituer l'image elle-même, ainsi que la différence avec la précédente. Une erreur relative à la différence réelle inférieure à 1.5% a été atteinte, avec un pixel photogrille de taille 15×15 µm² en technologie 0.5 µm, soit 30x30 λ² (équivalent à 3.9x3.9 µm² en 0.13 µm) pour une matrice de 256x256 pixels consommant 18 mW.

Enfin, nous avons vu lors de notre état de l'art sur les imageurs que les architectures de pixels actuelles à trois ou quatre transistors atteignent péniblement des dynamiques de l'ordre de 60dB. Certaines applications concernant des marchés potentiels très importants, le secteur automobile notamment, nécessitent des dynamiques proches de 120 dB. Longtemps, seuls les capteurs logarithmiques ont pu atteindre de telles dynamiques, avec l'inconvénient d'un bruit fixe trop élevé [Loose et al.-01] [Matou-03]. Aujourd'hui, des réalisations de type rétine ont permis grâce à un conditionnement spécifique au sein du capteur, d'atteindre des dynamiques de 110 dB [Benthien et al.-00], voir plus [Stoppa et al.-02].

IV.2.c. Quelques succès commerciaux

Chacune de ces réalisations représente une prouesse technologique, pourtant la plupart d'entre elles restent confinées au stade expérimental et n'ont pas été commercialisées. Une des raisons à cela est que le marché potentiel est souvent limité. Mais certains paris se sont tout de même avérés fort judicieux et lucratifs, grâce à leur aspect novateur, ce qui motive le monde de la R&D à une quête permanente de nouveaux dispositifs.

Dans le milieu des années 90, Mitsubishi Electric a réussi une production de masse de ses rétines artificielles et capteurs intelligents pour atteindre le nombre de 500 millions d'unités vendues en 1999 et un chiffre d'affaire de 128 milliards de dollars ! Les secteurs visés étaient alors les jeux, les systèmes de surveillance et sécurité, et le multimédia. Les fonctions intégrées par ces rétines étaient l'acquisition d'images, le fenêtrage, l'extraction de contour, la projection d'images (2D en 1D), et le filtrage [Kyuma-99]. Les avantages de ces capteurs CMOS, par rapport aux CCD d'alors, étaient une faible consommation, un coût réduit et une plus grande rapidité de traitement. Un exemple de rétine proposée comportait une matrice de 256x256 pixels, chacun de taille 17.5x13 λ² (soit 2.3x1.7 µm² en CMOS 0.13µm), pour un fonctionnement en mode intégration et des traitements en courant [Funatsu et al.-97].

La souris optique a elle aussi été un beau succès au début des années 90, date de la création de la société Logitech en Suisse, qui a commercialisé ce dispositif et fait encore parti aujourd'hui des leaders de ce marché. Un premier circuit avait été développé par [Tanner & Mead-86], basé sur la corrélation horizontale ou verticale de pixels voisins dont les luminances étaient supérieures ou inférieures à un certain seuil. La performance atteinte pour un rapport optique de 1 est une résolution de 100 pixels/pouce pour une vitesse maximale de 2.0 m/s. Ici aussi, chaque pixel fonctionnait en mode intégration et les traitements à partir de données en courant étaient mis en œuvre. La consommation excessive à cause de la nécessité d'une source lumineuse continue, qui constituait l'inconvénient majeur de ce circuit, a été adressée par [Arreguit et al.-96] pour obtenir un capteur de mouvement dont la résolution est de 800 pixels/pouce et une vitesse maximale de 0.3 m/s.

IV.3. Architectures et approche de conception « système »

Parmi ces diverses applications, il est important que ces capteurs soient des systèmes capables d'embarquer des traitements variés. Nous évoquons ci-après des rétines tesselations spécifiques, ainsi que des architectures programmables qui deviennent alors de véritables processeurs d'images au niveau pixel.

IV.3.a. Configurations géométriques

En effet, nous avons décrit au paragraphe II.1. les bases théoriques formalisant l'échantillonnage spatial dans les capteurs d'images, et en particulier pour les imageurs. Mais il faut savoir que certaines tessellations, ou maillages, présentent un intérêt reconnu pour quelques applications spécifiques. En effet, une disposition bien choisie des photorécepteurs sur le silicium peut permettre de ne s'intéresser qu'à un certain type de comportement spatial, réduisant ainsi dès l'acquisition les données à traiter aux seules informations effectivement utiles.

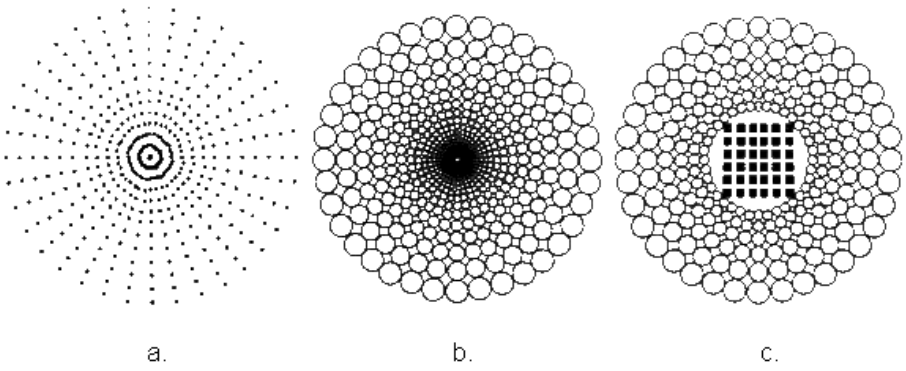


Figure I.29. Tessellations fovéales : a. polaire-linéaire, b. log-polaire, puis c. log-polaire avec centre cartesien.

Deux répartitions sont particulièrement intéressantes : la répartition hexagonale de la Figure I.11 (page 16) et les fovéales représentées ci-dessus sur la Figure I.29. Plusieurs capteurs dédiés, également appelés « rétines », ont été développés en mettant à profit ces répartitions particulières.

[Mahowald & Mead-89] puis [Andreou & Boahen-95] ont d'abord intégré une répartition hexagonale, par mimétisme avec la rétine des vertébrés. Le capteur résultant implémentait en effet un modèle incluant les photorécepteurs de la couche plexiforme externe (cones et bâtonnets), ainsi que leurs interactions (cellules horizontales et bipolaires). [Jolion-01] rappelle qu'un intérêt de cette répartition est une meilleure couverture des fréquences spatiales (économie d'environ 15%) par rapport à un maillage carré.

[Tremblay et al.-93] et [Tremblay et al.-95] ont, quant à eux, mis à profit l'intérêt majeur d'une structure hexagonale : l'équidistance entre pixels voisins dans les six directions principales de cette tessellation. En effet, cette dernière facilite grandement le traitement des données issues du capteur en prédisposant celui-ci à la tâche d'extraction de fronts de contrastes circulaires.

Le maillage fovéal (cf. Figure I.29) comporte une grande densité de détecteurs au centre, et une plus faible densité sur la périphérie. Il est particulièrement adapté aux opérations de zoom et de rotation, qui deviennent invariantes en coordonnées polaires, mais n'ont aucun intérêt en imagerie conventionnelle.

[Van der Spiegel et al.-89] furent les premiers à concevoir et fabriquer ce type de circuit, en employant une tessellation log-polaire et cartésienne au centre[33], avec une technologie CCD. [Wodnicki et al.-95], [Pardo et al.-97][Pardo et al.-98], et [Sandini et al.-00] ont, eux, réalisé quant à eux des intégrations en technologie CMOS, en ciblant notamment l'application de suivi d'objet. Effectivement ces systèmes à perception fovéale détectent une zone d'intérêt à l'aide des détecteurs périphériques, puis se focalisent sur celle-ci pour en acquérir les détails. Son atout majeur est de réaliser ce processus à complexité de calcul réduite, grâce au compromis judicieux trouvé entre résolution, champ visuel, et nombre de pixels à traiter.

La géométrie des photorécepteurs peut aussi jouer un rôle important quand il s'agit de prédisposer la photosensibilité à un certain type de comportement spatial.

IV.3.b. Rétines programmables

Au vue de la variété d'applications existantes et de l'intérêt certain du parallélisme massif et intrinsèque des capteurs d'images, notamment en termes de capacité de calcul. Certains auteurs ont développé des rétines à traitements intra-pixels programmables. [Paillet et al.-98] ont proposé une architecture SIMD[34] numérique 1-bit limitée à des opérations booléennes entre pixels. Des traitements numériques sur plusieurs bits ont également été conçus, par [Ishikawa et al.-99] par exemple.

Enfin, une autre approche regroupe des architectures analogiques programmables. Elles réalisent les opérations de calcul à l'aide de structures à capacités commutées [Dupret et al.-02], ou en mode courant [Rodriguez-Vazquez et al.-04] [Dudek & Hicks-05]. Nous avons inclus à titre indicatif en page Annexe D et E les caractéristiques du circuit de vision réalisé par ces derniers auteurs, ainsi que le schéma fonctionnel

---

[33] Cette configuration est ici aussi un mimétisme avec la biologie vivante, en particulier la rétine humaine ici, puisqu'elle est constituée en termes de densité de photorécepteurs de trois zones principales : une zone dense au centre, puis une relativement petite zone « vide », et enfin un zone moins dense en périphérie.

[34] Abréviation de « Single Instruction Multiple Data ».

d'une cellule de la matrice. Les traitements d'images possibles sont la convolution, le filtrage, la détection de contours, la détection et l'estimation du mouvement, le calcul d'histogrammes, ainsi que les opérations de morphologie mathématique.

Ces circuits sont caractérisés par une puissance de calcul de l'ordre de 10 MIPS par pixel, une précision de 7 bits environ, une consommation de l'ordre de 100 µW par pixel, et une taille pixel autour de 80 µm de côté ! A titre d'exemple, pour une opération de détection de contours à cadence de 25 im/s, un pixel consomme alors 13 nW.

Ces circuits de vision programmables s'avèrent particulièrement puissants mais leur surface pixel considérable vis-à-vis des pixels images rend impossible leur intégration à la résolution désormais courante de plus d'un million de pixels.

CONCLUSION

Dans ce chapitre, nous avons décrit un état de l'art sur les capteurs d'images CMOS. Nous avons commencé par les photodétecteurs existants en technologie CMOS ainsi que par quelques éléments d'optique et de photonique, et nous avons ensuite distingué deux familles de dispositifs : les imageurs et les rétines.

Nous avons présenté le conditionnement du signal photogénéré pour application en imagerie, les pixels « images », puis les paramètres qui caractérisent les imageurs. Nous avons explicité les principaux bruits influant sur leurs performances, pour terminer par une discussion sur le contexte industriel actuel. Concernant les rétines, nous avons développé le conditionnement du photosignal pour des applications nécessitant des traitements spatiaux puis temporels, les pixels « intelligents ». Nous avons alors présenté quelques exemples de réalisations de ces capteurs, afin de mieux définir les traitements envisageables au niveau pixel. Ceci dans la perspective de la définition de notre architecture de capteur.

En ce qui concerne les imageurs, nous avons relevé que la diminution de la taille des pixels entraîne une réduction des performances d'acquisition, ainsi qu'une limite en l'état actuel des technologies optiques autour de 2 µm de côté. Afin de continuer à réduire la taille des pixels, de nouveaux procédés de fabrication, ajoutant une nouvelle couche au dessus de celles de la technologie CMOS de manière à séparer la photodétection du substrat, sont en cours de développement. Il s'agit des technologies dites « Thin Film on Asic » TFA ou « Above IC ». Cependant elles ne sont pas encore complètement au point aujourd'hui.

Au sujet des rétines, la résolution atteinte de nos jours par les imageurs, qui dépasse le million de pixels, interdit l'intégration de traitements au niveau pixel. En effet, augmenter la taille d'un pixel engendre un coût en surface du capteur qui ne pourrait se justifier que par une valeur ajoutée très importante, et uniquement si la fonction réalisée ne peut pas s'obtenir autrement. Un traitement à ce niveau peut cependant être mis en oeuvre lorsqu'il y a nécessité de détecter, dès la photodétection, des variations lumineuses ou des zones particulières de la scène. On exploite alors le parallélisme intrinsèque des capteurs d'images, afin de réaliser des opérations de bas niveau avec une capacité de

calcul très importante, tout en réduisant la consommation et les ressources matérielles requises. En effet, les traitements de haut niveau seront alors limités à certains pixels de l'image seulement.

En revanche, un traitement de type digital et séquentiel sera plutôt préféré dans les cas où le traitement peut être intégré à une architecture numérique existante, donc potentiellement à moindre coût. Un interfaçage avec l'environnement extérieur, imposant généralement un protocole de communication digital pour assurer la transmission fidèle des données, peut aussi orienter le choix vers cette solution architecturale.

Notre objectif est d'ajouter aux imageurs fabriqués par la société STMicroelectronics la stabilisation vidéo. Cette fonction requiert la connaissance du mouvement entre les images acquises, aussi nous consacrons le chapitre suivant à cette tâche et à son intégration silicium.

Chapitre II.

ESTIMATION DU MOUVEMENT, THEORIE ET CAPTEURS

INTRODUCTION

Une grande partie du marché des imageurs est aujourd'hui portée par le développement des téléphones portables. Ce sont des appareils légers, petits, et donc très sujets aux tremblements et au « bougé ». Les vidéos acquises dans ces conditions sont particulièrement désagréables à regarder. Ce problème n'est pas neuf et il y a longtemps que les fabricants de caméscopes ont développé des solutions de stabilisation vidéo, qu'elles soient mécaniques ou électroniques.

Malheureusement la taille de ces applications portables interdit d'envisager une solution mécanique, à moins d'utiliser un MEMS. Les solutions électroniques sont beaucoup plus envisageables, mais elles requièrent une grande puissance de calcul, d'une part pour détecter le mouvement global de l'image, et d'autre part pour le compenser.

Il est encore trop tôt pour envisager de réaliser un imageur fournissant une vidéo stabilisée. En revanche, un constructeur d'imageurs CMOS tel que STMicroelectronics est prêt à embarquer sur un imageur des dispositifs qui permettront de faciliter la tâche de stabilisation. On parle alors de valeur ajoutée à un imageur. Une option qui n'introduit pas d'augmentation significative du coût du dispositif, mais qui peut représenter un avantage face à la concurrence.

Nous proposons donc ici d'étudier les solutions applicables à la réalisation de la fonction « estimation du mouvement global entre deux trames consécutives de la vidéo », sans dégrader la qualité de l'image acquise. En effet, il sera possible, à partir de ce mouvement global inter trames, de restituer une vidéo stable et fluide. Aussi ce chapitre II est-il dédié à la perception du mouvement.

Nous présentons tout d'abord le contexte de la stabilisation vidéo, qui nous permettra de définir dans le prochain chapitre les spécifications de notre capteur.

Nous abordons ensuite le contexte et les techniques d'estimation du mouvement, en présentant les performances, avantages et inconvénients de chacune, tout en gardant à l'esprit notre objectif d'intégration capteur, donc les ressources matérielles nécessaires et performances atteintes.

Enfin, nous présenterons un état de l'art sur les rétines, ce qui nous permettra par la suite de définir au mieux nos choix de l'architecture système de notre capteur.

I. LE CONTEXTE DE LA STABILISATION VIDEO

Deux familles de solutions existent pour stabiliser l'acquisition de séquences vidéo : l'une est de type mécanique et l'autre électronique.

I.1. Stabilisation mécanique

Les meilleures stabilisations vidéo s'obtiennent mécaniquement, à l'aide d'une optique ou d'un imageur mobile qui « filtre » les mouvements brusques du dispositif d'acquisition de manière physique. On agit dans ce cas directement au niveau du flux lumineux incident. Les mécanismes sont généralement constitués de senseurs gyroscopiques qui, via une électronique de commande, pilotent l'optique ou l'imageur mobile pour compenser les mouvements brusques transmis à l'appareil.

La Figure II.1 illustre cette correction optique dans le cas d'une optique commandée. Dans cette configuration, c'est la direction du flux lumineux incident qui est corrigé pour qu'il atteigne de manière stable et précise l'imageur.

Canon a été le premier fabricant à proposer ce type de technique et reste aujourd'hui un acteur majeur du domaine [Furukawa & Tajima-76] [Fujisaki et al.-01]. Par la suite, plusieurs sociétés comme Nikon, Konica Minolta, Panasonic ont elles aussi développé leurs solutions, avant que les fabricants de téléphones mobiles proposent à leur tour des dispositifs de stabilisation intégrés [Dutta et al.-03].

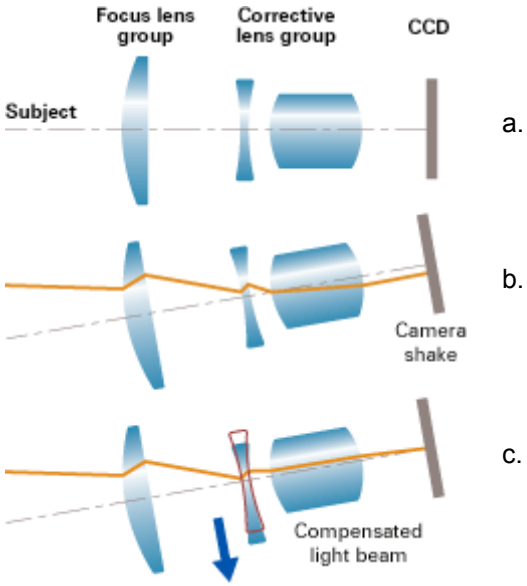


Figure II.1. Exemple de stabilisation mécanique par translation d'optique. Dans ce croquis[35], le flux lumineux de la scène visée n'est pas corrigé dans le cas b., alors qu'il l'est grâce à une lentille mobile et commandée dans le cas c.

[35] http://www.canon.com/technology/dv/02.html

Cette famille de stabilisateurs, en stabilisant directement le flux lumineux incident sur l'imageur, permet d'obtenir une qualité d'image optimale. En effet, pendant la période d'exposition des pixels à la lumière, chacun d'entre eux reçoit continûment et précisément la même zone de la scène visée, l'échantillonnage spatial résultant est alors d'excellente qualité. De plus, des essais montrent qu'une action mécanique sur l'optique plutôt que sur l'imageur fournit de meilleurs résultats [DigitalPhotography-06].

Cependant ces dispositifs mécano optiques restent relativement encombrants et coûteux, ce qui est rédhibitoire pour le marché des téléphones portables grand public que nous visons.

I.2. Stabilisation électronique

La stabilisation électronique constitue une alternative à cette approche de stabilisation vidéo [Morimotto & Chellappa-96] [Engelsberg & Schmidt-99]. Elle s'opère en post traitement des images provenant de l'imageur.

Comme illustré sur la Figure II.2, on peut caractériser la présence d'un bougé lors d'une prise de vidéo à l'aide de trois vecteurs mouvements : le mouvement global inter trame (ie. « mouvement effectif »), le mouvement que l'opérateur souhaite imprimer à la séquence (ie. « mouvement intentionnel ») et le mouvement perturbant (ie. « mouvement indésiré »).

Une vidéo stabilisée ne comporte idéalement que le mouvement intentionnel, la stabilisation consiste alors à compenser le mouvement indésiré en ne considérant qu'une fenêtre réduite des trames reçues (cf Figure II.3). Cette fenêtre est de taille fixe, correspondant à la résolution de l'affichage souhaité[36], et de position reconfigurée à chaque nouvelle image en fonction du mouvement indésiré.

La vidéo restituée et visualisée a posteriori est donc en réalité la séquence des fenêtres d'images, qui apparaît stable à l'utilisateur.

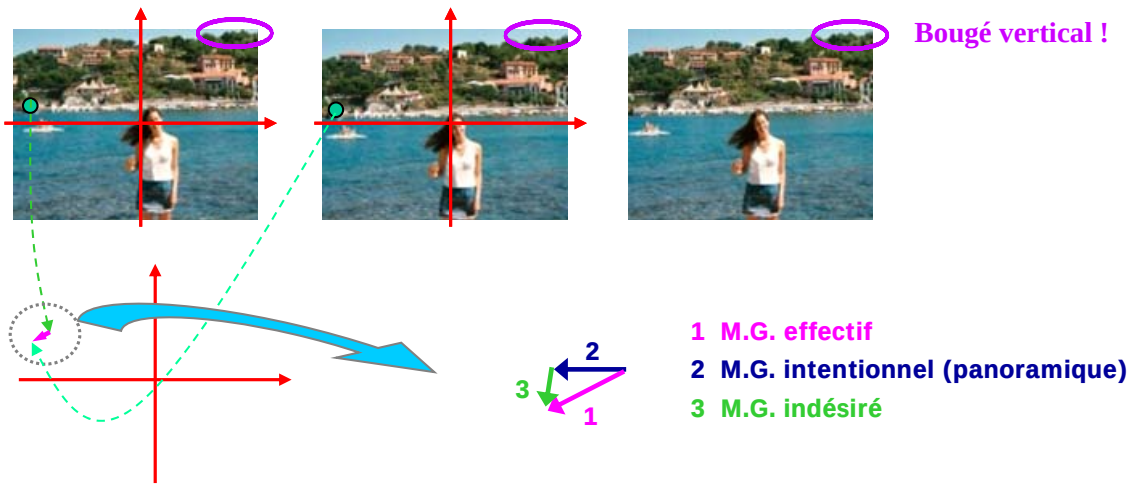


Figure II.2. Extrait d'une vidéo panoramique horizontale, avec bougé vertical lors de l'acquisition de la 2ème image.

[36] *Quelques formats standards sont : le VGA = 640×480, le SVGA = 800×600, le XGA=1024×768, SXGA = 1280×1024, le UXGA = 1600×1200.*

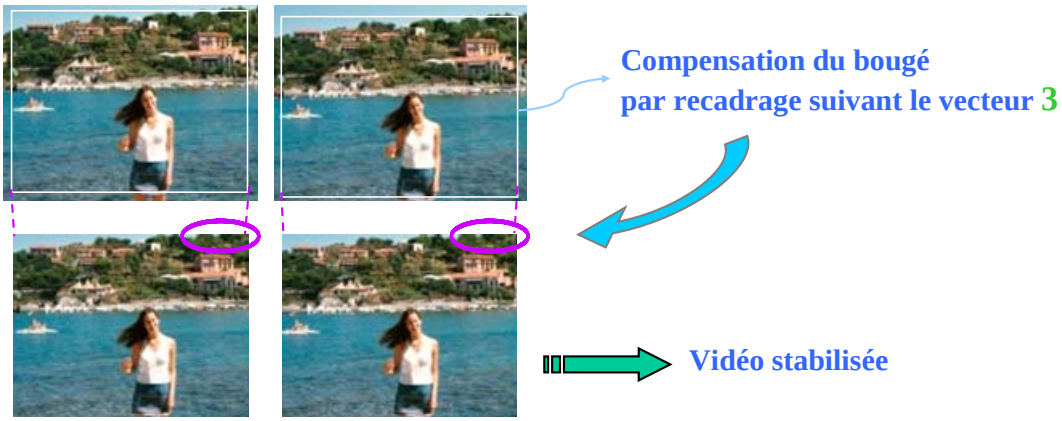


Figure II.3. Stabilisation d'une vidéo panoramique horizontale, avec bougé vertical. Le recadrage compense les variations brusques de mouvement global inter trames.

La tâche essentielle est donc le positionnement de cette fenêtre, c'est-à-dire la détermination du mouvement global inter images indésiré. Ce dernier est défini comme étant la différence du mouvement moyen, assimilé au mouvement intentionnel (un mouvement panoramique par exemple), et du mouvement global effectif.

La première étape en vue de la stabilisation consiste donc à estimer le mouvement global inter trames. Le mouvement intentionnel peut ensuite être déterminé par filtrage passe-bas temporel de ces mouvements globaux inter trames. Enfin le mouvement global indésiré, définissant le vecteur de recadrage, est obtenu par différence entre ces deux mouvements : effectif et intentionnel. Le diagramme de la Figure II.4 illustre cette procédure de stabilisation.

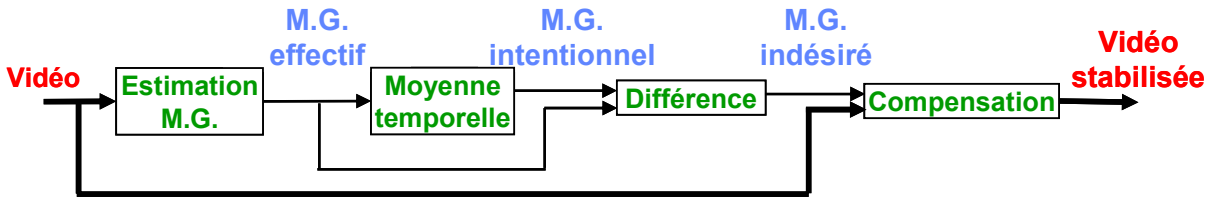


Figure II.4. Procédure de stabilisation vidéo numérique.

La pertinence du recadrage dépendra donc directement de la précision de l'estimation du mouvement global et du filtrage temporel permettant d'obtenir le mouvement intentionnel. [Comby-01] a montré que l'opération de filtrage temporel peut être réalisée par un filtre exponentiel, ce qui d'un point de vue matériel requiert très peu de ressources. Par contre, comme le montrent [Kuhn & Stechele-98] dans le cadre de la compression MPEG-4, la détermination du mouvement global inter trames constitue la tâche la plus complexe et nécessite beaucoup de ressources matérielles. [Auberger & Miro-05] proposent par exemple un algorithme de stabilisation vidéo de résolution CIF (320×240 pixels) implémenté en temps-réel vidéo (30 Hz) sur un processeur ARM 926EJ-S cadencé à 66 MHz. L'estimation du mouvement requiert dans ce cas 50 à 70 % des ressources nécessaires au processeur pour stabiliser la vidéo. Aussi, concentrons nous nos recherches sur cette tâche d'estimation du mouvement global.

Par ailleurs, le prix à payer pour cette technique de stabilisation vidéo est un nombre de pixels plus important que réellement restitué en sortie. Il s'agit donc d'une surface silicium ajoutée, donc un surcoût, que nous évaluons au chapitre III, et que nous discutons lors de notre analyse « d'Adéquation Algorithme Architecture Application » au Chapitre IV.

II. PERCEPTION ARTIFICIELLE DU MOUVEMENT

La mesure du mouvement apparent est exploitée dans tous les systèmes à base de vidéo numérique, où le déplacement des pixels au cours d'une vidéo est de grande importance.

Les standards de vidéo numériques par exemple sont pour la plupart basés sur la détermination de ces déplacements pixels pour réduire la quantité de données à transmettre. Tous les dispositifs de capture de vidéo ou d'images portatifs intègrent donc cette mesure du mouvement afin d'améliorer leur qualité d'image, de vidéo, ou de visiophonie.

Les accessoires informatiques tels que les souris optiques, les compresseurs MJPEG, ou encore les systèmes de vidéoconférence font aussi partie des applications qui en font grand usage, de même que les dispositifs de vidéo surveillance installés dans nos villes.

Cette mesure du mouvement est également essentielle en robotique pour asservir et piloter des systèmes mobiles.

L'inventaire d'applications ou de dispositifs basés sur la donnée du mouvement pourrait être très long, d'autant que de nouveaux systèmes visuels voient le jour en permanence. Nous assistons actuellement, par exemple, à l'avènement de dispositifs interactifs qui reposent sur l'analyse de la gestuelle des utilisateurs pour lancer des applications. La vidéo 3D fait, elle aussi, l'objet de grands intérêts et devrait connaître une forte croissance dans les années à venir.

Le problème de l'estimation et de la mesure du mouvement dans une vidéo mobilise par conséquent une large communauté scientifique depuis une vingtaine d'années. Des solutions algorithmiques et matérielles performantes ont été développées mais de nouveaux challenges apparaissent. Celui qui nous concerne ici est l'intégration de l'estimation du mouvement pour modules d'imagerie CMOS embarqués dans les téléphones mobiles. Si les ressources calculatoires ne permettaient pas il y a une vingtaine d'années de réaliser des tâches de traitement d'images en temps réel vidéo (25 images/s) [Koch-91], il en est autrement aujourd'hui et nous pouvons accomplir des tâches de vision de plus en plus complexes et de haut niveau.

II.1. Problématique

Le mouvement bidimensionnel dans une séquence vidéo résulte de la projection des mouvements des objets réels sur un capteur visuel. Cette projection d'une scène réelle 3D sur un espace 2D implique nécessairement une perte d'information spatiale contenue dans la scène, ce qui rend l'analyse de ces informations planaires compliquée. En effet à une variation de luminosité donnée ne correspond pas

forcément un phénomène physique donné. On se trouve alors en présence d'un problème dit « mal-posé »37 [Hadamard-1902][Bertero et al.-88].

Ceci se matérialise par exemple en considérant une scène fixe, sous éclairage mobile. Dans ce cas les objets sont immobiles et pourtant la séquence d'images acquise comporte des zones en mouvement : les ombres des objets.

L'estimation du mouvement dans une séquence d'images fait partie de ces analyses complexes. Elle consiste en l'estimation du mouvement 2D apparent de la scène projetée, appelé **flot optique** [Horn & Schunck-80]. Le flot optique est un champ de vecteurs mouvements dans la scène, comme illustré ci-dessous.

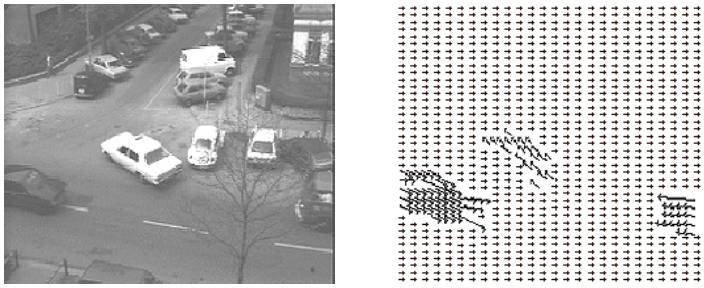


Figure II.5. Exemple d'une vidéo capturée avec un mouvement panoramique horizontal, avec le flot optique associé.

Dans cet exemple, nous pouvons distinguer deux sortes de mouvements : celui des objets à l'intérieur de la scène (ici, des voitures) et celui de l'ensemble de la scène (ici, un déplacement de gauche à droite). C'est ce dernier mouvement qui nous intéresse pour la stabilisation vidéo. Nous l'appellerons dorénavant « mouvement global ».

On peut noter que si ce mouvement est directement perçu sur l'arrière plan fixe de la scène filmée, il ne peut l'être sur des objets animés d'un mouvement propre. Ainsi, la manière la plus évidente d'estimer ce mouvement global est de mesurer le mouvement de l'arrière plan fixe de la scène.

Les variations d'éclairement de la scène dues au déplacement de la source lumineuse sont, quant à elles, plus problématiques pour notre cadre applicatif d'estimation du mouvement global pour la stabilisation vidéo. En effet elles vont générer une sensation de mouvement semblable à celle qui correspond à un déplacement de l'appareil d'acquisition. Ce type de variation ne sera cependant pas considéré ici car il est relativement peu fréquent.

D'autre part, tout système visuel a besoin, pour percevoir un mouvement, que la scène visualisée soit texturée. Nous illustrons ceci sur la Figure II.6 ci-dessous où nous avons séparé, sur un extrait de

---

[37] Un problème « mal-posé » signifie, au sens originel de [Hadamard-1902] qu'il peut ne pas avoir de solution unique, ou bien ne pas avoir de solution du tout…

séquence vidéo, la zone texturée de celle qui ne l'est pas. Nous voyons alors que le mouvement de translation horizontal inter images n'est perceptible que dans la zone texturée.

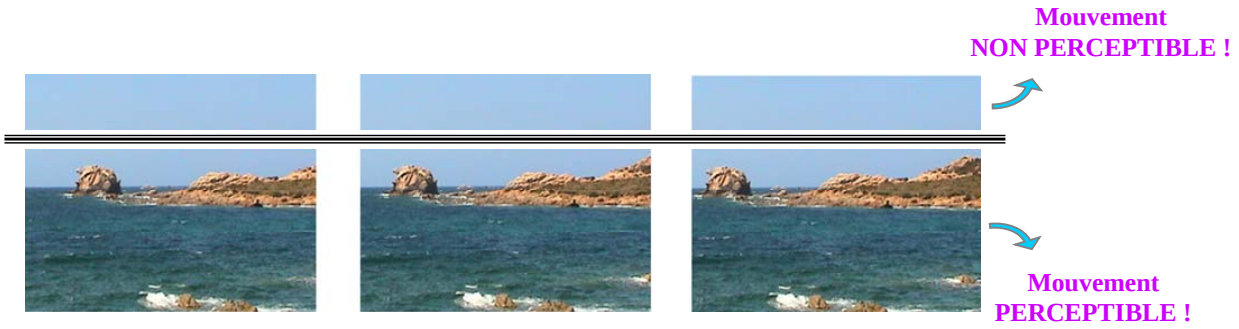


Figure II.6. Perception du mouvement et contraste visuel.

II.1.a. Le problème d'ouverture

Le *problème d'ouverture* est intrinsèque à tout système de perception visuelle [Hildreth-83], il ne s'agit pas d'un phénomène lié aux capteurs électroniques car les êtres vivants y sont également confrontés. Il est l'illustration du fait qu'une zone d'un objet dans une scène n'a pas une représentation unique sur le plan image [Ullman-79].

Le problème d'ouverture se caractérise par la perception du mouvement uniquement suivant la direction du gradient spatial d'une texture donnée. Ainsi, si dans le champ de vision considéré un objet est caractérisé par une texture qui ne possède qu'une seule direction de gradient spatial, le mouvement perçu ou mesuré sur l'objet sera la projection de son mouvement réel sur la direction du gradient (cf.Figure II.7).

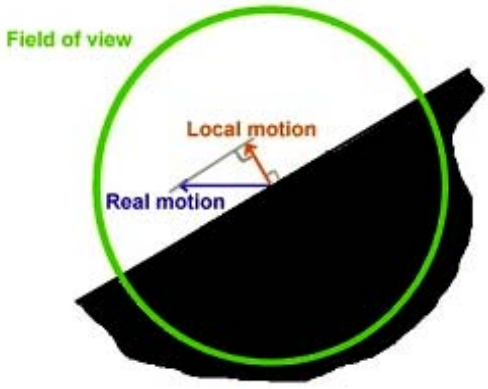


Figure II.7. Le problème d'ouverture.

Ce phénomène est omniprésent dans les systèmes de vision et se résout en associant les mouvements locaux dans au moins deux directions différentes de gradients spatiaux. Le corollaire est donc que le mouvement d'un objet ne peut être correctement perçu que si cet objet possède deux directions de gradients spatiaux différentes. La perception des mouvements d'une scène résulte donc de l'interprétation a posteriori de mesures locales et élémentaires du mouvement, appelées **flot optique**.

Chez l'homme par exemple, l'analyse visuelle du mouvement est obtenue par la succession de différents niveaux hiérarchiques constituées de couches de réseaux neuronaux [Jolion-01][Perrinet-02][Giese & Poggio-03].

Il est donc important de noter que pour estimer le mouvement d'un point d'une image, il faut tenir compte de son voisinage spatial ou spatio-temporel. Notre état de l'art des techniques d'estimation du mouvement qui est présenté dans les paragraphes II.2. et II.3. confirme ce constat.

II.1.b. Systèmes visuels biologiques et perception du mouvement

Les systèmes de vision biologique constituent depuis longtemps une source précieuse d'inspiration et d'idées pour le développement de systèmes artificiels [Franceschini-99].

La sélection naturelle agissant, la constitution des systèmes de vision biologiques qui existent aujourd'hui n'est pas que le fruit du hasard, elle est souvent optimale pour la situation concernée. Ainsi il nous semble important de nous intéresser à ces systèmes afin d'en bénéficier, le cas échéant, pour la conception de notre capteur. En effet, la difficulté à résoudre une tâche de vision donnée dépend fortement de la manière avec laquelle l'information lumineuse est acquise.

Des études sur la vision des animaux [SNOF] mettent ceci en évidence et montrent par exemple que la disposition des yeux et leurs structures sont étroitement liés à la recherche de nourriture. Ainsi le milieu dans lequel ils vivent, la rapidité avec laquelle ils se déplacent, le type de nourriture qu'ils doivent chasser et saisir, la vigilance dont ils doivent faire preuve pour échapper eux-mêmes à leurs prédateurs, sont autant d'éléments qui ont mené à des systèmes visuels parfois très différents.

La différence est frappante entre les rapaces qui possèdent une vision binoculaire[38] caractérisée par des yeux franchement orientés vers la même direction (cf. Figure II.8) afin de mieux percevoir une proie à grande distance, et les oiseaux communs à vision monoculaire qui ont leurs yeux orientés dans deux directions bien distinctes (cf. Figure II.8).

Figure II.8. Disposition des yeux chez un rapace (à gauche), et chez un rouge gorge (à droite).

[38] *Une vision binoculaire signifie que les champs de vision des deux yeux se superposent. A la différence de la vision monoculaire où les champs de vision sont distincts.*

Nous distinguons ici d'une part les concepts relatifs à une approche plutôt macroscopique de ces systèmes visuels, c'est-à-dire des considérations de disposition ou de géométrie des capteurs. Et d'autre part les solutions qui se rapprochent plus du niveau microscopique.

Considérations « macroscopiques ».

En remarquant que les êtres vivants qui volent ou nagent, donc qui évoluent dans un espace à trois dimensions, sont généralement dotés de systèmes à vision panoramique[39], [Fermuller & Aloimonos-98] ont démontré qu'il est plus simple d'estimer le mouvement en trois dimensions à l'aide d'un capteur sphérique que d'un capteur plan du type d'une camera caractérisé par un champ de vision réduit. Ce dit capteur sphérique peut être un œil ou tout système fournissant une vision panoramique.

Ces travaux ont notamment amené les auteurs à proposer deux nouvelles architectures de systèmes pour la perception 3D. Le premier dispositif est appelé « l'œil composé » (à gauche sur la Figure II.9), il rassemble des capteurs orientés vers l'extérieur autour d'une sphère. L'autre, dit à sphéricité « négative », (à droite sur la Figure II.9) est constitué de caméras captant l'information visuelle vers l'intérieur de la scène. Ces concepts ont ouvert la voie à diverses applications dans le domaine 3D en facilitant la perception dans cet espace. Les applications potentielles concernent essentiellement les systèmes 3D interactifs basés sur la reconnaissance de mouvement.

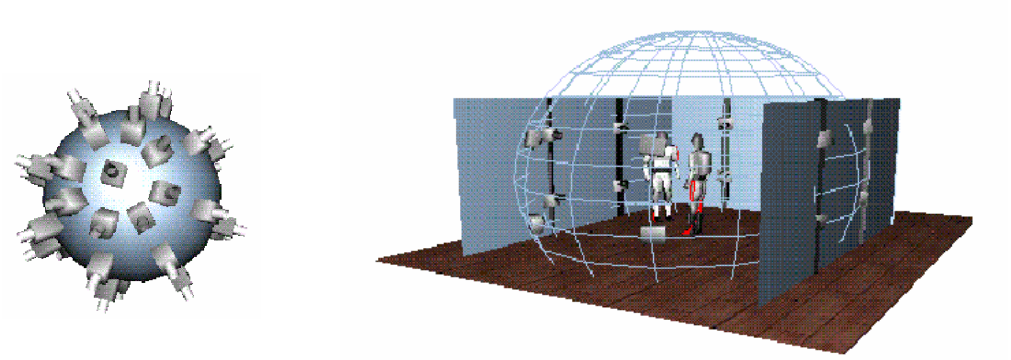


Figure II.9. Deux architectures d'acquisition visuelle sphériques dédiés à la perception 3D : « l'oeil composé » (à gauche) et le système à sphéricité « négative » (à droite).

D'autre part, comme nous l'avons évoqué au chapitre précédent, la manière dont est réalisé l'échantillonnage spatial du flux lumineux reçu peut faciliter certains traitements par rapport à d'autres. Ainsi une vision fovéale comme celle de l'homme a l'avantage de limiter le nombre d'information à traiter tout en restant capable de voir ce qui nous intéresse en haute résolution. Nous réussissons cela en ciblant la région centrale sur le point intéressant pour en percevoir les détails, les photodétecteurs périphériques permettant de détecter ce point d'intérêt à moindre coût de traitement. Ceci est un exemple d'utilisation astucieuse de l'échantillonnage spatial.

[39] *Ces systèmes sont généralement des arrangements de deux capteurs visuels orientés dans des directions très différentes afin d'élargir le champ de vision.*

Considérations « microscopiques ».

La perception du mouvement a été étudiée par les neurobiologistes qui en ont proposé des modèles. De nombreuses expériences menées sur l'homme et le singe ont par exemple permis d'établir que cette perception est le résultat d'un traitement hiérarchique dans le cerveau qui, à partir de mesures très localisées et élémentaires permet d'aboutir à la reconnaissance de mouvements plus globaux et complexes. [Giese & Poggio-03] proposent un modèle basé sur l'hypothèse que la représentation du mouvement dépend de schémas « appris »[40] par rapport aux formes d'objets de la scène et aux déplacements de régions singulières dans la scène, ces dernières étant caractérisées par un champ spécifique de vecteurs mouvements locaux, comme illustré sur la Figure II.10 ci-dessous.

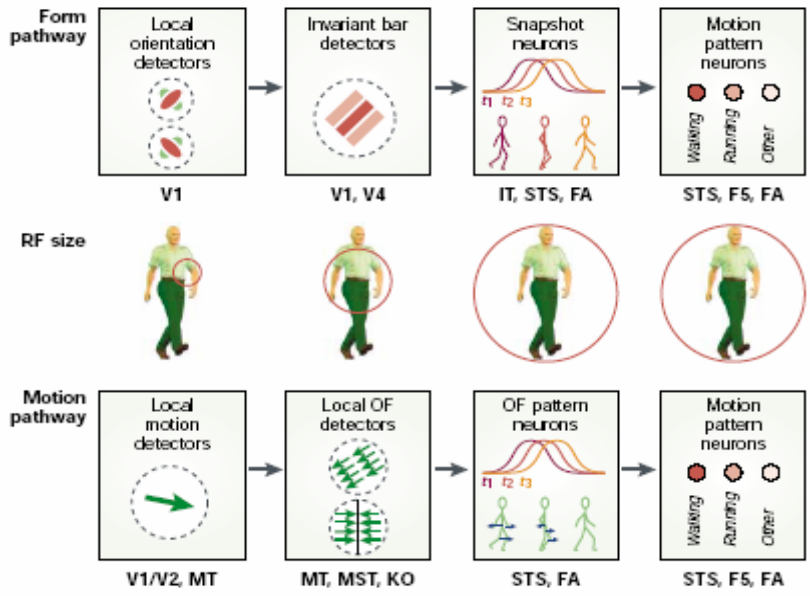


Figure II.10. Modélisation hiérarchique du traitement pour la perception des formes (en haut) et du mouvement (en bas) [Giese & Poggio-03].

Le modèle résumé dans cette figure montre bien cette hiérarchie de traitements et décrit bien la croissance de la complexité des évènements détectés au fur et à mesure de la progression du traitement du signal dans la hiérarchie. Les auteurs distinguent le traitement des formes et du mouvement sur cette figure, tout en sachant que les deux interagissent. Chaque niveau se caractérise par des architectures spécifiques de traitement, nous nous intéressons ici particulièrement au mouvement (en bas sur la Figure II.10).

Le premier niveau est celui des phototransducteurs, le cortex visuel primaire « V1 ». Il rassemble les détecteurs locaux du mouvement sélectifs à une direction donnée. Il existe différents modèles pour ces détecteurs [Borst & Egelhaaf-89] [Grossberg et al.-00]. Ils ont la propriété commune d'implémenter

[40] *Les traitements visuels sont en effet obtenus à partir de réseaux de neurones, pour lesquels l'apprentissage est une notion intrinsèque.*

localement un filtrage spatio-temporel du signal lumineux[41]. [Adelson & Bergen-85] ont montré l'équivalence de nombreux modèles à celui initialement proposé par [Reichardt-61]. Ce dernier est essentiellement un corrélateur entre deux signaux provenant de photorécepteurs voisins (cf. Figure II.11). La direction préférée est celle de l'alignement des deux photorécepteurs et l'amplitude de la vitesse détectée dépend du déphasage imprimé au signal.

Grâce à la faisabilité de son intégration, ce principe de détection est à l'origine de plusieurs circuits et systèmes de détection du mouvement implémentés sur silicium [Andreou et al.-91] [Harrison & Koch-00] [Liu-00].

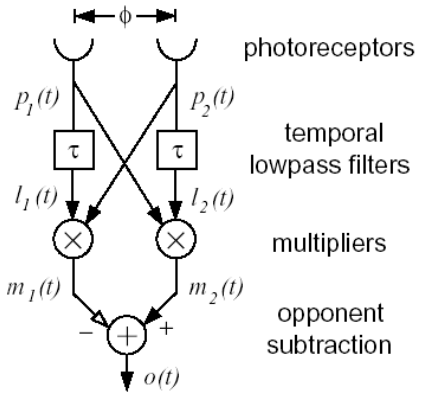


Figure II.11. Modèle de détection locale du mouvement de Reichardt [Reichardt-61].

Un modèle de ce traitement spatio-temporel rétinien est présenté dans [Jolion-01]. Il décrit l'existence de cellules dites « bipolaires », dédiées à la perception du contraste[42]. Ces cellules ou photorécepteurs réalisent un filtrage spatial de type passe-haut pour extraire ces contrastes. Ce filtrage est réalisé à partir de la différence entre le photosignal d'un photodétecteur et la moyenne des photosignaux voisins.
[Jolion-01] met en avant deux autres propriétés essentielles à la perception du mouvement. Il s'agit du comportement auto-adaptatif, permettant au photorécepteur de fonctionner sur une grande dynamique de lumière ambiante, et d'une plus grande sensibilité aux variations rapides de lumière, ce qui privilégie les zones contrastées de la scène. Ces deux aspects avaient été également remarqués par [Werblin-74], ce qui a motivé la conception par [Delbruck & Mead-89] d'une architecture de pixel auto-adaptative très performante (dixit [Koch & Mathur-96]). Celle-ci est mise à profit dans les rétines artificielles de perception du mouvement [Higgins et al.-05] [Stocker-06], mais aussi dans des rétines plus généralistes dédiées à l'étude des fonctions visuelles en neurosciences [Delbruck & Liu-04].

Le second niveau de traitement est constitué de récepteurs dont la fonction est d'analyser la structure des données provenant du niveau précédent (cf. Figure II.10). Nous pouvons classer ces récepteurs en

[41] *Le filtrage spatio-temporel permet d'extraire un mouvement dans une direction et à une vitesse préférées.*

[42] *Là aussi, un mouvement ne pourra être perçu que si la scène est texturée.*

deux catégories suivant leur fonction. Les uns sont sensibles au mouvement dans quatre directions préférentiellement et les autres segmentent les mouvements perçus dans les directions horizontales et verticales[43]. Ce distinguo des différent types de mouvements apparaît comme une tâche essentielle qui est présente dès la rétine chez le lapin et la salamandre et qui est suspectée chez beaucoup d'autres êtres vivants [Olveczky et al.-03].

Le troisième niveau est encore mal connu. Il pourrait être le lieu de détection des régions singulières caractérisées par un champ spécifique de vecteurs mouvements locaux[44]. Ces régions sont ensuite interprétées dans le quatrième niveau, par filtrage temporel, pour privilégier la détection de séquences d'évènements.

La notion d'intégration, en espace et en temps, est essentielle en perception du mouvement car elle permet d'obtenir une quantité suffisante d'informations spatio-temporelles pour déterminer le mouvement. [Burr-80] a montré que le système visuel humain procédait à une intégration des informations de type microscopique sur une durée voisine de 100ms.

II.2. Estimation locale du mouvement, le flot optique

Nous avons vu précédemment que le problème d'ouverture impose de considérer un voisinage spatial ou spatio-temporel pour déterminer le mouvement local de manière juste. Nous présentons ici les principales techniques qui ont été proposées pour estimer ces mouvements locaux constituant le flot optique de la vidéo. Elles peuvent être classées en trois catégories :

- ➙ les méthodes différentielles,
- ➙ les méthodes de mise en correspondance explicite,
- ➙ les méthodes fréquentielles de corrélations spatio-temporelles.

Pour tout point réel (X,Y,Z,T) de la scène, ayant une vitesse V et se projetant sur l'image en (x,y,t), le flot optique informe d'un mouvement v en (x,y), formé des deux composantes orthogonales dx et dy.

Couramment, l'estimation de mouvement est basée sur l'hypothèse de conservation de la quantité de lumière de tous les points de l'image. Ainsi, le pixel projeté en (x,y) voit sa luminance se conserver :

Eq. II.1. $$I(x,y,t) = I(x+dx, y+dy, t+dt)$$

Cette formule exprime le déplacement du pixel caractérisé par son éclairement I(x,y,t).

II.2.a. Les méthodes différentielles

L'équation du flot optique modélise le mouvement d'un pixel dans une séquence d'images. Les paramètres spatiaux et temporels sont alors pris en compte. Cette équation repose sur l'hypothèse

[43] *On pourra se référer à [Giese & Poggio-03] pour plus d'informations et références.*

[44] *La constitution de ce niveau est encore mal connue.*

d'illumination constante de l'image acquise. Elle provient donc de l'équation Eq. II.1, en considérant une fréquence d'échantillonnage élevée [Horn & Schunk-80]. Pour résoudre cette équation, on la développe en série de Taylor - de type : $y(t+\delta t) = y(t) + \delta t.\frac{dy}{dt} + \frac{\delta t^2}{2}\frac{d^2y}{dt^2} + ...$ -, on obtient (au premier ordre) :

Eq. II.2. $$I(x+dx, y+dy, t+dt) = I(x,y,t) + \frac{\delta I}{\delta x}dx + \frac{\delta I}{\delta y}dy + \frac{\delta I}{\delta t}dt + R$$

R est le reste de Taylor représentant les ordres 2 et supérieurs des dérivées. En utilisant Eq. II.2 dans Eq. II.1 et en négligeant le reste R, on obtient l'équation :

Eq. II.3. $$\frac{\delta I}{\delta x}dx + \frac{\delta I}{\delta y}dy + \frac{\delta I}{\delta t}dt = 0$$

En divisant par dt et en posant que $\frac{dx}{dt} = u$ et $\frac{dy}{dt} = v$, on obtient l'équation du flot optique :

Eq. II.4. $$\frac{\delta I}{\delta x}u + \frac{\delta I}{\delta y}v + \frac{\delta I}{\delta t} = 0$$

$\frac{\delta I}{\delta x}$ et $\frac{\delta I}{\delta y}$ sont les variations spatiales (gradients), $\frac{\delta I}{\delta t}$ exprime les variations de la luminosité dans le temps. Cette équation donne la relation du mouvement formé des 2 composantes orthogonales u et v, et d'une illumination restant constante dans un domaine tridimensionnel (x,y,t : spatio-temporel), tel que l'intensité d'un pixel (x,y,t) donné se retrouve un instant dt plus tard à une autre position (x',y',t').

On appelle cette équation « l'**équation de contrainte du mouvement** » (ECM), et on peut aussi l'écrire sous la forme :

Eq. II.5. $$\overrightarrow{grad}\, I.\vec{V} + \frac{\partial I}{\partial t} = 0$$

La donnée seule de l'ECM ne suffit pas pour obtenir une solution unique, c'est le problème d'ouverture. Une hypothèse supplémentaire sur la forme du déplacement que l'on souhaite trouver doit alors être ajoutée pour obtenir cette solution, c'est la régularisation.

Différents types de régularisations ont été proposées et comparées dans [Barron et al.-94]. Le premier travail réalisé par [Horn & Schunck-80] consiste à choisir une régularisation quadratique globale. Le problème de l'estimation du flot optique s'écrit alors sous la forme d'une minimisation de la somme des carrés des dérivées premières spatiales en x et en y, avec une contrainte de lissage globale[45]. Mais les résultats montrent un lissage trop important et indésirable aux niveaux des discontinuités du mouvement (qui sont a priori inconnues). Afin d'éviter ce problème causé par le choix d'un lissage dit « isotropique »,

[45] *C'est-à-dire sur toute l'image.*

des lissages « anisotropiques » ont été envisagés [Cohen & Herlin-93] [Aubert et al.-99], ainsi que « anisotropiques et temporels » [Weickert & Schnörr-01].

Une autre approche pour contraindre le problème consiste à considérer que le mouvement est uniforme et de type translation sur un voisinage local de l'image, c'est l'hypothèse proposée par [Lucas & Kanade-81]. Le mouvement est alors contraint par un lissage sur un voisinage local. Comme le montrent [Barron et al.-94], cette approche est la plus performante et constitue de plus un bon compromis entre charge de calculs et performances [Baker & Matthews-04].

D'autre part, pour un même échantillonnage temporel, l'équation de contrainte du mouvement n'est plus valide dans le cas de mouvements de grandes amplitudes et une approche multi échelle est alors mise en œuvre [Burt & Adelson-83] [Barron et al.-94]. C'est-à-dire que l'on construit une pyramide d'images (Gaussienne par exemple) pour appliquer la technique d'estimation du mouvement sur la couche de plus haut niveau tout d'abord, puis sur le niveau inférieur, jusqu'à l'estimer sur l'image originale. L'atout principal de cette approche est d'alléger la charge de calcul.

II.2.b. Les méthodes de mise en correspondance

L'approche de mise en correspondance met explicitement en évidence l'hypothèse de base en recherchant directement une zone d'une image dans la suivante. Il faut alors trouver le déplacement qui apparie au mieux deux régions ou éléments caractéristiques (contour ou coin par exemple) de la première image dans la suivante. L'appariement est alors effectué en minimisant un indice de dissemblance.

Les techniques les plus communes sont appelées « block matching », elles consistent à reconnaître un bloc de pixels d'une image dans la suivante[46]. Bien qu'imprécises à cause de la discrétisation du déplacement estimé (le pas est le pixel), ces techniques possèdent les qualités essentielles de robustesse, de simplicité de mise en œuvre, d'estimation de déplacements de grandes amplitudes[47] et de bon compromis entre précision du mouvement et quantité de données transmises [Dufaux & Moscheni-96]. Elles se retrouvent dans la majorité des standards de compression vidéo, notamment ceux de la norme « mpeg », la prédiction du contenu des blocs s'obtient à partir du mouvement estimé [Golston-04] [Andersson et al.-02].

Deux blocs de pixels sont appariés s'ils minimisent un critère de dissemblance ou une distance donnée. L'équation ci-dessous exprime cette recherche de minimum dans une zone de recherche « S », pour des blocs de taille « W ».

---

[46] En codage vidéo, la taille des blocs est communément de 8×8 ou 16×16 pixels. Aujourd'hui, des techniques apparaissent où la taille des blocs est variable afin d'améliorer la qualité de la compression, notamment aux frontières des objets en mouvement [ACIVS'05].

[47] A la différence des techniques différentielles qui, en se basant sur les dérivées partielles de la luminance, peuvent ne pas aboutir en cas de bruit trop important, ou d'un nombre réduit d'images disponibles (espace mémoire limité).

Eq. II.6.
$$d = \min_{\vec{d} \in S} \sum_{\vec{r} \in W} \left\| I(\vec{r}, t) - I(\vec{r} + \vec{d}, t + \Delta t) \right\|$$

L'indice de dissemblance à minimiser constitue un point critique et déterminant pour obtenir un appariement pertinent.

Deux métriques sont largement utilisées en compression vidéo : l'une est la somme des différences absolues[48] alors que l'autre est la somme des carrés des différences[49] [Dufaux & Moscheni-95].

La mise en correspondance des blocs peut être menée de manière exhaustive. Dans ce cas on obtient l'estimation du mouvement optimum, mais au prix d'une charge de calcul importante (de l'ordre de 50 kop. pour chaque vecteur mouvement estimé). Afin d'améliorer le compromis entre performances et coût calculatoire, des techniques hiérarchiques de recherche ont été mise en œuvre comme l'algorithme de recherche en trois étapes, dit « three-step search »[50] [Koga et al.-81] [Song & Chun-04].

Par ailleurs, afin de tenir compte du fait que l'estimation du déplacement des blocs de pixels est plus difficile à obtenir sur des zones peu texturées, l'estimation du déplacement d'un bloc de pixel tient compte de son contenu. C'est le cas pour le standard MPEG-2 par exemple [Andersson et al.-02].

II.2.c. Les méthodes de corrélation

Le principe fondamental de conservation de la luminance d'un point au cours de son déplacement est implicite pour les techniques de corrélation. Ces méthodes sont issues des études sur les systèmes de vision biologiques [Adelson & Bergen-85], elles se formalisent mathématiquement comme suit :

Soit une image décrite par sa composante lumineuse I(x,y,t), la transformée de Fourier associée peut s'écrire $\hat{I}(f_x,f_y,f_t)$. Soit r_x et r_y les composantes horizontales et verticales d'un vecteur vitesse, la transformée de Fourier de l'image en mouvement est donnée par :

Eq. II.7 $$TF(I(x-r_x t, y-r_y t, t) = \hat{I}(u,v,w+r_x u+r_y v)$$

Les fréquences spatiales sont inchangées, mais toutes les fréquences temporelles sont translatées par le produit de la vitesse et des fréquences spatiales. Il s'agit alors d'identifier dans l'espace des fréquences un plan de vitesse d'équation $w+r_x u+r_y v = 0$ pour retrouver les composantes r_x et r_y associées au déplacement.

L'information de mouvement est en général extraite par des filtres orientés spatialement et temporellement (filtres de Gabor par exemple). On distingue deux approches de résolution : l'une basée sur l'énergie du signal [Heeger-88] [Spinei et al.-98], l'autre sur la phase [Fleet & Jepson-89].

---

[48] *« SAD », pour « Sum of Absolute Differences ».*

[49] *« SSD », « Sum Squared Differences ».*

[50] *Ces techniques se basent sur l'hypothèse que l'indice de dissemblance augmente de manière monotone lorsque l'on s'éloigne du minimum global.*

II.3. Estimation du mouvement global

Le flot optique est une image, d'une part due au mouvement des objets mobiles dans la scène, et d'autre part due au déplacement de la caméra. Ce dernier mouvement nous intéresse en vue de stabiliser la vidéo.

Outre la stabilisation vidéo, la donnée de ce mouvement global inter images sert diverses applications.

➙ La compression vidéo, où l'information du mouvement global est exploitée pour réduire la quantité de données « image » à transmettre pour décrire une séquence vidéo (standard MPEG-4 [Golston-04] [BroadWare-05]). Dans le cas des scènes où le déplacement des pixels d'une image à une autre peut être prédit, on ne transmet alors que le vecteur mouvement de la zone « image » (de l'ordre de quelques octets) plutôt que de transmettre tous les pixels constitutifs de la zone déplacée.

➙ La « super résolution », qui permet d'augmenter la résolution des images [Elad & Feuer-97]. Son principe consiste à estimer, à partir des images basses résolutions de départ et de la donnée précise du mouvement global inter images (précision sub-pixellique), la luminance des pixels correspondant aux images haute résolution désirées.

➙ La mosaïque d'images, qui consiste à construire une image de grandes dimensions à partir d'une séquence vidéo qui explore une scène [Lu et al.-03].

➙ La reconnaissance de mouvements prédéfinis imprimés par un opérateur à un dispositif d'acquisition portatif. Dans un cadre d'interface homme-machine, on peut imaginer que cette reconnaissance commande certaines fonctions d'un téléphone mobile, où participe aux éléments perceptifs d'un système interactif par exemples.

➙ la segmentation d'images au sens du mouvement, qui s'obtient en détectant les zones de l'image qui possèdent un mouvement différent du mouvement de la caméra. Cette information du mouvement global est par exemple utile en surveillance vidéo pour détecter des intrusions sur la zone à surveiller. Mais aussi en robotique mobile, où les zones en mouvements singuliers dans une scène fixe sont des obstacles, qui doivent être pris en compte pour pouvoir évoluer dans cet environnement.

Le mouvement global entre deux images consécutives est décrit par un modèle. Ce modèle résulte d'une analyse géométrique du mouvement, et sous quelques hypothèses sur la scène à percevoir (considérée quasi-plane, constituée d'objets rigides[51] et statiques), il se limite à quelques paramètres [Stiller & Konrad-99]. Ce sont ces paramètres que nous devons quantifier et qui permettent de décrire très efficacement le déplacement des pixels entre images consécutives (cf. Figure II.12).

Ainsi, les choix majeurs à effectuer pour réaliser une estimation du mouvement global concernent : le modèle du mouvement et la technique d'estimation des paramètres du modèle du mouvement.

[51] *C'est-à-dire des objets qui ne se déforment pas, à géométrie fixe.*

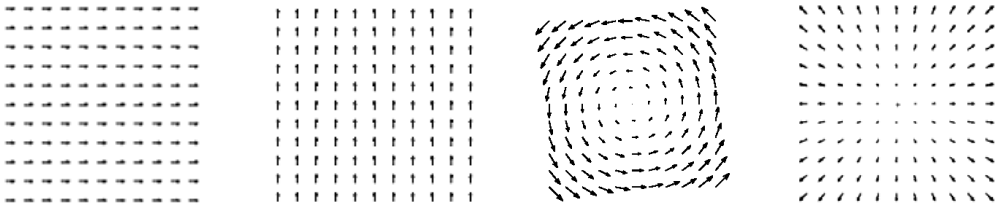


Figure II.12. Exemples de mouvements globaux décrits par un modèle à 4 paramètres.

II.3.a. Choix du modèle de mouvement

Le modèle du mouvement global à estimer est déterminant. Il va permettre de décrire plus ou moins fidèlement le mouvement global entre deux images consécutives, engendré par le mouvement de la caméra, et perçu au niveau du plan focal (image).

Deux modèles doivent être considérés : le modèle spatial et le modèle temporel.

- **Modèle spatial.**

Cinq modèles sont couramment utilisés pour décrire les mouvements inter images d'une séquence :

➙ Modèle 1 : translation (Tx,Ty), il s'agit d'un modèle à 2 paramètres, décrivant le déplacement en translation horizontale (Tx) et verticale (Ty) entre deux images.

➙ Modèle 2 : translation + rotation (Tx,Ty,θ), ici trois paramètres permettent de rendre compte des translations dans le plan image, ainsi que des rotations autour de la normale au plan image.

➙ Modèle 3 : translation + rotation + zoom (Tx,Ty, α.cos θ, α.sin θ), ce modèle considère 4 paramètres de mouvement pour permettre d'estimer, en plus des translations et rotations dans le plan image, le facteur de zoom α (ou facteur d'échelle) correspondant à une translation suivant la normale au plan image.

➙ Modèle 4 : translation + rotation + zoom + cisaillement + obliquité (Tx,Ty,a1,a2,a3,a4), six paramètres composent le modèle « affine ». Les mouvements de cisaillement et d'obliquité dus aux rotations autour des deux autres axes que la normale au plan image sont pris en compte.

➙ Modèle 5 : tout type de mouvement global, appelé modèle de « perspective », c'est un modèle qui se réduit à 8 paramètres lorsque l'on considère les trois hypothèses ''raisonnables'' suivantes :

- le mouvement de translation de la caméra reste faible comparé à la distance la séparant des objets de la scène,
- la scène est considérée approximativement plane,
- le champ de vision est faible.

Ces modèles sont donc plus ou moins complets, et nous savons que plus le modèle est complet moins l'estimation des paramètres sera stable car plus il a de chance d'être biaisé par des mouvements parasites dans l'image [Odobez-Bouthemy95].

En pratique, il est préférable de procéder de manière progressive en considérant tout d'abord le modèle 1 qui décrit les mouvements globaux les plus courants dans les séquences d'images[52], les translations. Puis la description du mouvement est affinée en complexifiant le modèle. L'estimation du modèle 1 représente généralement un très bon compromis entre complexité, donc charge de calcul, et validité du mouvement décrit [Hager & Belhumeur-98].

A noter que ces modèles permettent une mise en équation linéaire du problème, pour laquelle la résolution au sens des moindres carrés est possible et dont la charge de calcul est réduite, donc attractive pour l'application temps réel embarquée qui nous concerne.

- **Modèle temporel.**

Ce modèle définit l'évolution temporelle du modèle spatial. Il peut être soit linéaire, si on considère que le mouvement entre les deux images ne possède pas d'accélération ; soit quadratique, dans le cas contraire.

On se contente communément d'un modèle linéaire lorsqu'il s'agit d'estimer le mouvement entre deux images consécutives.

II.3.b. Choix du support d'estimation

Le support d'estimation constitue l'ensemble des pixels de l'image qui sont mis en jeu dans l'estimation des paramètres du mouvement, c'est-à-dire qui participent directement à l'estimation. Ce support doit être choisi avec attention car les performances de l'estimation en dépendent.

En effet, imaginons un système dans lequel on doit estimer le mouvement de rotation autour de l'axe optique[53]. Si l'on choisit d'extraire ce mouvement de rotation uniquement à partir des pixels très voisins du centre de rotation, le pas d'échantillonnage des données (le pixel) fera que la précision de l'angle estimé sera faible. En revanche, cette estimation de l'angle de rotation sera d'autant meilleure que les pixels impliqués seront éloignés de l'axe. Ainsi, le choix du support d'observation détermine les performances envisageables.

Une conséquence immédiate est que, si le support d'observation est mal choisi, même la meilleure des techniques d'estimation des paramètres du mouvement[54] ne pourra donner de bons résultats.

Quelques supports d'estimation :

- L'image entière.

---

[52] Nous considérons ici des séquences d'images de la vie courante, puisque acquises avec un dispositif du type téléphone portable.

[53] *C'est-à-dire l'angle de rotation autour de la normale centrée du plan image.*

[54] *Si elle existe…*

Les techniques d'estimation du mouvement global les plus robustes considèrent l'ensemble des données disponibles, c'est-à-dire l'image entière, pour contraindre au mieux les paramètres du modèle de mouvement à estimer [Odobez & Bouthemy-95]. La charge calculatoire est cependant très importante, liée au nombre de pixels pris en compte.

D'autre part, toute technique d'estimation du mouvement global doit être robuste à des éléments parasites éventuels, inconnus à priori, comme des objets mobiles dans la scène par exemple. Une démarche de segmentation au sens du mouvement est alors mise en œuvre afin de dissocier ces mouvements singuliers du mouvement global engendré par le déplacement de la caméra.

La charge de calculs associée est d'autant plus grande que cette robustesse doit être appliquée à l'ensemble de l'image.

Certaines techniques considèrent un sous-échantillonnage de l'image pour réduire la charge de calcul sans que les performances soient trop affectées. Le rapport « performances / charge de calculs » est ainsi augmenté.

Ce sous-échantillonnage correspond aux emplacements d'éléments caractéristiques des images[55] dans le cas des travaux de [Morimoto & Chellappa-96] [Censi et al.-99]. Il peut aussi être celui des vecteurs mouvements correspondant aux appariements de blocs de pixels caractéristiques de la compression vidéo [Tan et al.-00] [Smolic et al.-00].

- Des zones d'intérêts.

Plutôt que de considérer toute l'image, certains auteurs ne considèrent que quelques zones de celle-ci pour extraire le mouvement global, réduisant ainsi la charge calculatoire. [Uomori et al.-90] décomposent l'image en quatre zones identiques et rectangulaires (cf. Figure II.13). [Engelsberg & Schmidt-99] emploient une connaissance a priori sur le contenu de la scène en supposant que la scène filmée contient une personne filmée au milieu de l'image, l'estimation du mouvement global est alors réalisée en considérant trois zones identiques et rectangulaires, comme illustré sur la Figure II.13. Les auteurs précisent que seulement 5 à 10% des pixels de la zone d'intérêt suffisent pour obtenir une précision suffisante d'estimation, bien que cela dépende de la texture de la scène.

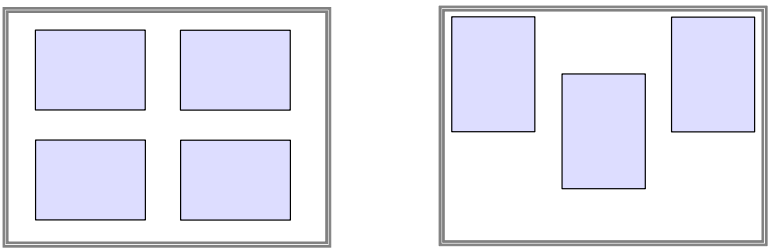


Figure II.13. Exemples de zones d'intérêts dans l'image proposées par [Uomori et al.-90] à gauche, et [Engelsberg & Schmidt-99] à droite.

[55] *Ces éléments sont typiquement des zones contrastées de l'image.*

II.3.c. Techniques d'estimation du mouvement global

Le mouvement global est décrit par un modèle qui contient peu de paramètres par rapport au nombre de mesures, nous avons donc affaire à un système surcontraint. Du fait des erreurs de mesure, un tel système ne possède généralement pas de solution.

La démarche pour répondre à cette situation se décompose en deux étapes. Tout d'abord on rend ce système sous contraint, en ajoutant des inconnues qui sont classiquement les erreurs entre le modèle et les mesures. Une infinité de solutions existent alors. Il nous reste à déterminer la solution unique en optimisant ce système sous contrainte. Cette contrainte est typiquement la minimisation des erreurs sus introduites au regard d'une métrique donnée.

L'algorithme de [Odobez & Bouthemy-95] par exemple met en oeuvre les statistiques robustes, plus précisément les M-estimateurs, dans un processus de résolution incrémental et multi échelle. Il a été optimisé et implémenté sur ultra-Sparc-I pour atteindre une vitesse de 1.6 estimations par seconde en pleine résolution et 2.3 estimations par secondes en arrêtant l'estimation à l'avant dernier niveau de la pyramide multi-résolution.

III. PERCEPTION DU MOUVEMENT AU NIVEAU PIXEL

Les traitements mis en jeu dans la majorité des systèmes qui restituent le mouvement s'opèrent en post-traitement des images acquises. Le traitement du signal vidéo numérisé est alors mené par une architecture numérique telle qu'un co-processeur, un DSP, ou un FPGA par exemple.

Or, il est possible de traiter l'information lumineuse au plus près de son acquisition, au niveau pixel, pour percevoir ce mouvement apparent. Nous décrivons dans cette section un état de l'art sur ces réalisations rétiniennes.

Avant cela, il nous paraît important de préciser les différences entre ces deux approches de traitement de l'information.

En post-traitement, outre l'échantillonnage spatial inévitable, l'image est échantillonnée temporellement. Du point de vue de la charge de calcul et de la surface silicium totale du système, ce n'est pas la meilleure approche pour une opération de détection telle que la perception du mouvement. En effet, l'incertitude sur la direction et l'amplitude de déplacement des pixels entre les deux instants d'échantillonnage, implique de mettre en œuvre des techniques de calculs complexes permettant de reconnaître un pixel ou une zone de la première image dans la deuxième.

La tâche est alors plus complexe que dans le cas d'un traitement en temps continu où il est possible, par exemple, de détecter le passage d'un élément caractéristique d'une image (front de contraste) d'un pixel à son voisin. Ainsi, le déplacement est connu puisque défini comme le passage d'un pixel à un autre, et il ne reste alors qu'à mesurer l'intervalle de temps correspondant pour obtenir la vitesse de

déplacement du front. En considérant cette vitesse comme constante pendant l'intervalle de temps entre deux images consécutives, on pourra obtenir le mouvement local inter images.

III.1. Mesures locales

Au cours de ce paragraphe, nous allons présenter l'état de l'art concernant les capteurs du mouvement local. Pour cela, nous avons conservé la classification adoptée dans la section précédente.

Ici, phototransduction et mesure du mouvement sont associées au sein même du pixel. Un mouvement local est directement mesuré à l'aide de pixels ''mouvement'' optimisés pour cette tâche. Cette approche possède des qualités de perception visuelle efficace, mais elle ne permet pas d'atteindre une grande densité de pixels. Pour une même surface silicium et un même coût, l'échantillonnage spatial du flux lumineux incident est alors réduit, et la qualité de l'image est donc dégradée. Un compromis entre densité et fonctionnalité du pixel est nécessaire.

D'un point de vue spatial, il est important de rappeler que le pas d'échantillonnage détermine la fréquence spatiale maximale à ne pas dépasser pour éviter les problèmes de repliement de spectre, et donc les fausses mesures. Dans certains cas, une légère défocalisation de l'optique permet d'adapter, dans une certaine mesure, le contenu spatial projeté sur le silicium à celui de l'échantillonnage du capteur.

Par ailleurs, le mouvement ne peut être perçu et quantifié que sur des zones texturées d'une scène. Ainsi, détecter ces zones avant de procéder à l'estimation du mouvement augmente la robustesse de la mesure.

III.1.a. Mise en correspondance d'éléments caractéristiques

Cette approche consiste à détecter un élément caractéristique dans la scène à un instant donné, puis de le reconnaître un instant plus tard sur un pixel voisin. On peut alors en déduire la direction du mouvement, et si on est capable de mesurer l'intervalle de temps entre les deux passages, on en déduit la vitesse de déplacement et donc le déplacement correspondant pendant une durée inter images.

Dans une approche logicielle du traitement des images acquises, cet élément à apparier peut être plus ou moins complexe : un front de contraste, un coin, ou un objet par exemple. Mais dans le cadre d'une implémentation dans le plan focal, où les contraintes technologiques en termes de surface silicium sont fortes, seules des caractéristiques peu complexes peuvent être extraites. Il s'agira généralement d'apparier des fronts de contrastes qui peuvent être détectés soit de manière temporelle, soit spatiale. Dans le premier cas, cela consiste à extraire une variation rapide de luminosité sur un même pixel [Delbruck-04] [Kramer-02] [Delbruck & Mead-96], et dans le deuxième à détecter une variation conséquente de luminosité d'un pixel à un autre [Vittoz & Arreguit-93] [Andreou & Boahen-95].

Le champ de vecteurs mouvements obtenu est épars[56], dépendant des caractéristiques spatiales de l'image perçue. Une texture pauvre en contrastes et motifs impliquera un nombre réduit d'éléments

[56] *Par opposition à un champ dense, où chaque pixel possède un vecteur mouvement.*

caractéristiques, donc peu d'information de mouvement. Par contre, ces éléments sont appariés de manière robuste et leurs vecteurs mouvements associés sont donc fiables. A l'inverse, si la scène est trop texturée, de faux appariements sont possibles : c'est le problème des appariements multiples. Il est dû à l'incertitude d'apparier le « bon » élément caractéristique à cause du sous-échantillonnage de l'information contenue dans la scène.

- Mise en correspondance temporelle

Alors que le traitement des images acquises tente d'extraire le mouvement à partir de données échantillonnées à intervalles de temps fixes, le traitement dans le plan focal en temps continu amène à mesurer le temps pour une distance spatiale fixe (entre deux pixels).

[Horiuchi et al.-91] s'inspirent du modèle biologique de perception du mouvement de [Barlow & Levick-65], et d'une architecture biologique de perception auditive pour déterminer le temps de passage de ce front en une position voisine [Konishi-86]. Il s'agit de détecter le passage d'un front de contraste positif (d'intensité croissante) entre deux pixels voisins. Chacun d'eux signale ce passage par une impulsion de durée fixe, ces deux impulsions se propagent alors en sens inverse sur deux lignes à retard, qui comportent des éléments de corrélation de type « ET » logiques, dont les sorties deviennent les entrées de circuits « winner take all » (WTA). La sortie valide du WTA indique alors le délai correspondant, donc la vitesse de passage de l'élément caractéristique entre les deux pixels. Pour une longueur d'impulsion donnée, le circuit peut mesurer des vitesses sur une gamme d'une décade, et pour différents réglages de la durée d'impulsion il a permis de relever des vitesses comprises entre 2.9 pixels/s et 1150 pixels/s. Le front montant de contraste[57] est détecté de manière temporelle grâce au photorécepteur adaptatif de [Delbruck-89] qui permet de détecter les variations rapides de luminosité et les corrélations sont effectuées en courant. De plus, un indice de confiance dans la mesure de la vitesse peut être obtenu par somme de courants. Grâce à la bonne sensibilité aux variations lumineuses du photorécepteur adaptatif, le circuit possède une réponse relativement invariante au contraste d'entrée. Cependant son exposition en lumière modulée (néons de 100 Hz, par exemple) le rend inutilisable en l'état car, tous les photorécepteurs réagissant en même temps, le circuit indique constamment une vitesse infinie.

[Kramer et al.-96][58] proposent eux aussi de mettre en œuvre le photorécepteur de Delbruck pour détecter un front de contraste de manière temporelle, puis de corréler les réponses entre deux pixels voisins pour en déduire la vitesse de passage. La mesure de mouvement est alors explicite pour les trois versions de capteurs proposées.

La première version, constituée de deux et appelée « facilitate-and-sample (FS) », est une structure qui initie une impulsion de tension au passage d'un front de contraste, qui décroît ensuite de façon

[57] *Ce front peut être choisi indifféremment montant ou descendant.*

[58] *Ces travaux ont été menés à Caltech (California institute of technology). Plusieurs brevets ont été déposés par rapport à ces travaux : [US5781648], [US6023521], [US6088467], [US6212288], [US6212289], et [US5998780].*

logarithmique dans le temps jusqu'à ce qu'une impulsion sur un pixel voisin échantillonne ce signal. La tension résultante est alors la représentation logarithmique de la vitesse de passage du front sur les deux pixels. La structure se caractérise par conséquent par une large plage dynamique de mesure de vitesses. Des mesures à l'aide de stimuli générés électroniquement ont effectivement attesté d'un bon fonctionnement sur près de 7 décades et sur 3 décades dans un cadre réel intégrant l'optique et sous lumière incandescente de bureau. Un autre avantage de taille est sa sortie sous la forme d'un échantillon de tension correspondant à la dernière mesure. En effet cette tension peut être conservée et lue de façon synchrone par les circuits de lecture d'un imageur par exemple, ce qui n'est pas directement le cas pour les deux autres structures. Le circuit est quasi-insensible aux variations d'amplitude du front de contraste si sa pente est rapide. Par contre, la vitesse mesurée à tendance à diminuer lorsque la pente du front de contraste diminue, c'est à dire lorsque le contraste est moins saillant (observation d'une différence de 10% sur la mesure). La taille du pixel fabriqué en lithographie CMOS 2µm est voisine de 220×220 µm, c'est-à-dire $110 \times 110\ \lambda^2$ soit encore 14×14 µm² en CMOS 130 nm.

Une seconde version, nommée cette fois « facilitate-and-trigger (FT) », comprend deux photodétecteurs. Elle consiste à générer une impulsion binaire de durée fixe lors du passage d'un front de contraste sur un pixel, puis une même impulsion est générée sur un pixel voisin. Un circuit extrait alors la partie commune de ces deux impulsions binaires (fonction équivalente à un ET logique). Cette durée d'impulsion constitue le temps de passage d'un pixel à un autre, donc la vitesse de déplacement du front puisque l'espacement inter photodétecteurs est connu. La vitesse est alors inversement proportionnelle à cette durée. Des mesures attestent le même comportement que la structure précédente en termes de sensibilité au contraste et dépendance à la pente du front. Par ailleurs, les auteurs soulignent la fidélité des mesures de vitesses sur une gamme d'intensités lumineuses ambiantes supérieure à 3 décades. La dynamique de vitesse, pouvant être mesurée, est ici aussi limitée à 1.5 décades environ.

Enfin, la version « facilitate-trigger-inhibit (FTI)», [Kramer-96], est une structure où trois photodétecteurs sont mis en œuvre pour percevoir la direction et l'amplitude du déplacement d'un front de contraste. Le premier détectant le front, il rend alors le deuxième sensible au passage éventuel du front. Si cela est le cas, ce deuxième photodétecteur initie une impulsion binaire dont la durée dépend de l'instant de passage du front sur le troisième photodétecteur qui va alors inhiber cette impulsion. Ainsi la vitesse mesurable n'est pas limitée puisque ici la durée de l'impulsion n'est pas fixée a priori, contrairement au cas précédent. Là encore la vitesse est mesurée de façon inversement proportionnelle à la durée de l'impulsion. La dynamique de mesure est par conséquent large en théorie, mais s'est avérée limitée à 1.5 décade environ en pratique. Comme dans le cas des deux structures précédentes, une campagne de mesures a montré que la structure était sensible à la pente du front de contraste, avec une diminution typique proche de 10% pour des fronts progressifs par rapport à des fronts abrupts. Enfin le circuit se caractérise par une insensibilité à l'intensité lumineuse ambiante sur une plage dynamique supérieure à 3 décades. La taille d'un pixel fabriqué en précision lithographique 2 µm est de 166×166 µm², donc $83 \times 83\ \lambda^2$, ce qui est de l'ordre de 10.8×10.8 µm² en CMOS 130 nm.

Ces trois derniers capteurs élémentaires de vitesse dans la plan focal ont été mis à profit avec succès dans un cadre de robotique mobile pour la localisation de points de divergence, du temps avant impact, de segmentation au sens du mouvement.

[Etienne-Cummings-93] et [Etienne-Cummings-01] proposent une autre technique de mesure de la vitesse de déplacement des fronts de contraste dans la scène. Elle repose sur la perception spatiale des contrastes saillants dans la scène puis la mesure du temps de passage de ceux-ci d'un pixel à un autre. Le front de contraste est détecté par filtrage spatial passe-haut réalisé à l'aide de réseaux résistifs et de circuits amplificateurs, avant d'être binarisé par comparaison au seuil avec hystérésis. Le circuit réalise ensuite la corrélation temporelle des fronts de contrastes d'un pixel à son voisin. En effet, la mise en correspondance n'est effective que si l'apparition d'un front est consécutive à sa disparition sur un détecteur voisin. La durée de l'impulsion résultante, correspondant au temps de passage entre ces deux pixels, pilote la charge d'une capacité à courant constant grâce à un amplificateur intégrateur. La tension obtenue indique alors la vitesse, et sa polarité la direction du déplacement. La cellule complète de mesure du mouvement local occupe une surface de 110×220 µm² en CMOS 2µm, soit 55×110 λ², c'est-à-dire de l'ordre 7,15×14,3 µm² en CMOS 130nm. Les vitesses des mouvements mesurées s'étalent sur deux décades environ, pour des luminosités ambiantes allant d'une faible luminosité intérieure (25 mW/m²) jusqu'à un éclairement extérieur ensoleillé (2,5 W/m²), et un contraste minimum requis de 20%.

Des travaux de Toyota R&D, [Yamada & Soga-03], traduisent électriquement le temps de parcours d'un front de contraste entre deux pixels. Les contrastes sont détectés en comparant non pas la différence des photocourants de deux photodétecteurs voisins, mais en comparant leur rapport. L'avantage de ces travaux est d'obtenir ainsi une grandeur indépendante de la luminosité ambiante. Le traitement étant en temps continu, les circuits logiques de mise en correspondance des impulsions initiées au passage d'un contraste sont asynchrones. Une campagne de mesures a montré une dynamique de mesures des vitesses de 2 décades pour une gamme d'intensité de lumière ambiante variant de 3 décades. La précision des mesures a été évaluée à ±20% de la vitesse réelle. La taille des pixels est de 570×135 µm² en CMOS 1.5µm, ce qui correspond à 380×90 λ². Ces dispositifs ont été employés pour la surveillance du trafic routier, où les vitesses détectées s'étalaient de 0.3 à 180 km/h. Ils permettent également le contrôle visuel de l'angle mort des véhicules, en détectant un véhicule en approche ou un piéton par exemple.

[Miller & Barrows-99] ont développé un capteur 1D de suivi de zones contrastées. La photoréception est logarithmique et les contrastes recherchés correspondent à un masque prédéfini ({-1,0,1} ou {-1,2,-1} par exemple) appliqués directement aux phototensions. Les sorties sont connectées à des structures « winner-take-all » qui identifient les plus forts contrastes et génèrent une impulsion. Des circuits mettent alors en correspondance ces impulsions et la vitesse peut être mesurée par temps de passage. La preuve du fonctionnement des pixels a été tout d'abord fournie à l'aide de motifs noir et blancs, avant d'être utilisés en tant que capteurs visuels pour la navigation de véhicules aériens autonomes [Green et al.-04].

Jusqu'ici nous avons mesuré le signe et l'amplitude de la vitesse dans une direction donnée[59]. Certaines réalisations se sont limitées à la perception du sens de déplacement du front de contraste afin d'être utilisées en robotique pour fournir un champ de directions des mouvements dans la scène et en extraire le point de divergence.

Dans ce but et dans la continuité des travaux réalisés à Caltech, [Higgins et al.-99] ont proposé deux variantes de pixels : « Inhibit-Trigger-Inhibit (ITI) » et « Facilitate-Trigger-Compare (FTC) ».

La première version initie une impulsion binaire qui est inhibée par l'un des pixels voisins. Seul ce voisin est alors actif et génère un potentiel « vitesse » valide. La sortie du capteur élémentaire est un courant fonction de la différence des deux tensions « vitesses » détectées dans les deux directions possibles. Ce courant indique alors directement la direction du mouvement. Le pixel complet contient trois photodétecteurs, deux unités de détections du mouvement, et un étage de sortie. Il occupe en lithographie 1.2µm une surface de 110×120 µm², soit 92×100 λ² en surface relative.

La version FTC se compose de deux photorécepteurs. Lorsqu'un front traverse l'un d'eux, le signal de facilitation est alors activé pendant un temps fixe. Si ce front traverse le pixel voisin pendant cet intervalle de temps, une impulsion est générée par ce dernier. L'étage de sortie compare les potentiels de ces deux pixels et génère un courant constant positif si le mouvement est vers la droite, négatif si il est vers la gauche. Le pixel résultant est de taille 128×119 µm² en CMOS 1.2 µm, c'est à dire 107×100 λ².

Ces deux circuits ont des performances très semblables, avec une dynamique de détection de la direction du mouvement de 2 décades pour un contraste minimum de 40%.

[Ruedi-96] a également proposé un pixel de détection du mouvement fonctionnant sur ce même principe de corrélation temporelle par temps de passage d'un front de contraste détecté cette fois de façon spatiale. L'auteur effectue la différence du photocourant d'un photodétecteur avec celui traversant un réseau résistif de transistors en faible inversion (filtrage passe-bas) ainsi qu'une comparaison à hystérésis. Dans le cas de la détection d'un front, l'auteur génère alors deux impulsions de durées distinctes que l'on corrèle avec celles initiées par le pixel voisin afin d'être sensible dans 2 gammes de vitesses. Le circuit est sensible à une vitesse minimum de 3 pixels par seconde et jusqu'à 25 pixels par seconde. De plus un maillage hexagonal est mis en œuvre pour détecter les mouvements dans 3 directions, chaque 120°. La taille du pixel est de 223×215 µm² en CMOS 2µm, soit 111×107 λ².

- Mise en correspondance spatiale

Il est également possible de mettre en correspondance des pixels ou des éléments caractéristiques d'une scène, non pas entre deux positions spatiales données, mais entre deux instants donnés. C'est la démarche classiquement employée lorsque l'on doit traiter la succession d'images acquises par vidéo. Il faut donc reconnaître un élément de la scène échantillonné à l'instant t1 (image 1), à l'instant t2 (image 2).

[59] *Etant donné que localement le mouvement d'un objet est perçu perpendiculaire à son bords (problème de l'ouverture), la mesure de la vitesse suivant une direction donnée est en réalité la projection du vecteur vitesse sur cet axe.*

Des pixels spécifiques peuvent extraire ces éléments spatiaux qui devront être « reconnaissable » entre ces deux instants. C'est-à-dire qu'ils doivent être autant que possible insensibles aux perturbations lumineuses. Les zones contrastées de la scène sont typiquement robustes à cela.

L'intérêt de détecter ces zones au niveau pixel est de réduire la quantité de données à traiter ensuite, en se limitant à celles-ci, donc de réduire les ressources nécessaires et la consommation du système. Dans notre cas, ces informations extraites au niveau pixel doivent être acquises en même temps que les images afin de fournir les données les plus concordantes possibles avec la vidéo pour sa stabilisation.

[Moini et al.-93] [Moini et al.-97] réalisent un codage spatial de l'image en considérant l'évolution temporelle de la luminance de chaque pixel, qui est codée sur trois états : luminance croissante, décroissante, ou stable. A partir de ce codage spatial, on recherche, en post-traitement, les différents codes d'une image transformée à une autre. Un seuil sur la dérivée temporelle du photosignal est utilisé pour décider de la croissance, décroissance, ou stabilité de la luminance. Les photodétecteurs sont des photodiodes de substrat avec évolution logarithmique de la photo tension grâce à un transistor connecté en diode.

L'opération de dérivée temporelle se caractérise par une influence indésirable lors d'une utilisation en lumière modulée car elle amplifie ces variations, et tous les pixels réagissent en conséquence à ces fluctuations. [Moini et al.-97] remédient à ce bruit multiplicatif en divisant le signal de chaque pixel par la moyenne du signal des voisins. Le rapport signal à bruit est alors amélioré de 20 dB environ. Chaque pixel est de taille environ 130×130 λ^2, soit 16×16 μm^2 en process CMOS 130nm.

[Park et al.-03] extraient par approche bio-mimétique les fronts de contrastes en temps continu au niveau pixel, par réseau résistif de transistor en fonctionnement de faible inversion. Les zones contrastées obtenues sont alors mises en correspondance d'un instant à un autre par post-traitement sur FPGA.

Par ailleurs, [Ruedi et al.-03] ont mis en œuvre un capteur d'extraction de la direction et l'amplitude du contraste qui encode de manière asynchrone et par contrastes décroissants les positions des pixels correspondants. Le capteur est capable de mesurer des contrastes minimums de 2%, leur orientation de 0 à 360° avec une précision de ±2°, et ceci sur une dynamique de lumière dans la même scène s'étalant sur 120dB. La taille d'un pixel fabriqué en CMOS 0.5µm est de 69×69 μm^2, c'est-à-dire 138×138 λ^2. La perception du mouvement des zones contrastées dans la scène pourrait alors être menée en mettant en correspondance les positions des contrastes entre deux instants.

Enfin, les travaux de thèse de David Navarro menés ici au LIRMM ont consisté à réaliser un codage spatial de l'image, afin d'obtenir, pour chaque pixel un code binaire fonction de sa texture avoisinante [Navarro-03]. Ce code comporte en réalité le signe du gradient spatial existant entre le pixel central et chacun de ses voisins, c'est-à-dire 8 bits pour ses huit voisins. Le mouvement d'un pixel de l'image I1 est ensuite obtenu en recherchant sa nouvelle position dans l'image suivante I2. Cette mise en correspondance est là aussi réalisée en post-traitement, sur FPGA. Chaque pixel occupe une surface de 40×50 μm^2 pour un procédés CMOS 0.35µm, c'est-à-dire une taille relative de 114×142 λ^2. Les caractérisations des circuits fabriqués ont permis de rendre compte d'une trop faible robustesse de codage.

En effet, pour deux scènes statiques et acquises de façon identique, 40% des codes sont différents d'une image à l'autre. Une trop grande sensibilité au bruit du comparateur est à l'origine de ce phénomène.

III.1.b. Approche différentielle

Nous avons vu que le contenu visuel spatio-temporel d'une scène peut être extrait en résolvant l'équation de contrainte du mouvement (Eq II.3. p 15). Si les solutions de cette équation renseignent bien sur les mouvements dans les images, elles ne fournissent qu'une idée de ceux-ci et ne sont que qualitatives. Il parait donc à priori difficile de les mettre en œuvre pour stabiliser une vidéo.

L'un des premiers systèmes intégrés sur silicium implémentant et résolvant en temps continu cette équation est celui de [Tanner & Mead-86]. Une information spatio-temporelle représentative du mouvement global de la scène perçue est reportée à l'aide de deux tensions, chacune renseignant une composante du vecteur mouvement global. Celle-ci est obtenue en minimisant la somme des carrés des erreurs de l'équation de contrainte du mouvement grâce à un réseau de photodétecteurs interconnectés mesurant les gradients spatiaux et temporels locaux en chaque nœud. La donnée « vitesse » est alors issue du rapport de ces gradients spatiaux et temporels. Ce circuit de 8×8 pixels s'est montré très sensible aux variations lumineuses et au contraste, même dans le cadre de sources contrôlées. La surface silicium requise en fabrication CMOS 1.5 µm est de 4500×3500 µm².

[Deutschmann & Koch-98a] ont montré que l'équation de contrainte du mouvement pouvait effectivement être résolue avec une telle approche. Le circuit est une implémentation 1D de cette approche différentielle. Les caractérisations montrent un bon fonctionnement du dispositif, avec notamment un comportement robuste aux variations lumineuses et une bonne linéarité. Ce circuit permettrait de détecter des vitesses s'étendant sur 2 décades pour une taille de pixel de 147×270 µm² en CMOS 2 µm, soit 73.5×135 λ².

Les mêmes auteurs, [Deutschmann & Koch-98b], ont implémenté une version 2D de ce type de circuit. Une matrice 15×15 a notamment été conçue, sur laquelle le mouvement est rendu compte par produit des gradients spatiaux et temporels de la scène. Le circuit est très dépendant du contraste et du spectre fréquentiel et spatial de la scène, cependant il décrit très fidèlement la direction du mouvement. Chacun des pixels occupe avec une technologie de 1.2 µm une surface de 112×112 µm², correspondant à 93×93 λ².

[Stocker-06] présente la réalisation la plus aboutie de ces circuits percevant le mouvement. Chaque pixel fait partie d'un réseau de pixels interconnectés, chacun fonctionnant en tant que circuit rétroactionné qui compare ses mesures locales de gradients spatio-temporels avec une estimation de la vitesse globale d'un voisinage de pixels, pour ensuite corriger cette estimation pour réduire la différence perçue. Le comportement du circuit est indépendant du contraste dès lors qu'il est supérieur à 30%. La taille pixel est de 124×124 µm² en CMOS 0.8µm, donc 155×155 λ².

III.1.c. Corrélations spatio-temporelles

La dernière famille de techniques de perception du mouvement a fait l'objet de plusieurs réalisations, souvent à caractère bio-inspiré. Ces réalisations, comme dans le cas précédent, ne rendent pas compte de la perception dans le plan focal des vitesses réelles de déplacement des éléments de la scène. Il s'agit plutôt d'une indication d'un contenu spatio-temporel.

[Benson & Delbruck-91] ont développé une rétine capable de percevoir le mouvement dans une direction et autour d'une lumière ambiante données. Cette rétine implémente le modèle biologique de perception du mouvement du lapin [Barlow & Levick-65], favorisant la détection d'un front de contraste dans une direction donnée. Le circuit intégré comporte une matrice de 41×47 pixels pour une surface silicium de 4.6×6.8 mm² en précision lithographique 2 µm. Au passage d'un front de contraste, un pixel génère une impulsion dont la durée dépend de son contraste et de sa vitesse. Cette information ne code donc pas explicitement la vitesse de passage du front.

Plusieurs travaux ont concerné le modèle de perception de Reichardt. Une première réalisation de [Andreou et al.-91] consiste en une matrice 1D de pixels au sein desquels la multiplication du signal photoélectrique filtré spatialement, par la version retardée d'un pixel voisin, réalise la fonction de corrélation. Une moyenne des courants de sortie de chacun des pixels est alors réalisée pour obtenir une information globale du mouvement vers la droite et la gauche, le mouvement final en étant la différence. Les auteurs ont mesuré des vitesses comprises entre 0 et 160 pixels/s.

La première implémentation 2D revient à [Delbruck-93], qui a utilisé des lignes à retards 1D comme filtres temporels pour détecter les fronts de contrastes. Ces lignes à retards sont orientées suivant trois directions spatiales, sous forme d'une tessellation hexagonale et perçoivent les déplacements lumineux suivant une direction et une vitesse données. Les variations spatio-temporelles perçues sont donc pseudo-locales. Le photosignal ainsi généré est intégré temporellement. Chaque pixel mesure 225 µm de côté en technologie 2 µm, soit une surface relative de 112.5 λ².

[Harrison & Koch-00] et [Liu-00] ont implémenté eux aussi le modèle de perception de Reichardt. Le mouvement visuel est perçu en multipliant le photosignal après filtrage temporel passe-haut d'un pixel, par le photosignal de sortie de filtres temporels passe-haut et passe-bas d'un pixel voisin. Le dimensionnement des constantes de temps des filtres passe-haut et passe-bas font que ces pixels sont très sélectifs à certaines fréquences spatiales et temporelles du signal lumineux. Dans le cas du circuit de [Harrison & Koch-00], la taille des pixels fabriqués avec une précision de 1.2µm est de 61×199 µm², c'est-à-dire 50×165 λ².

[Liu-00] propose une agrégation spatiale de ces perceptions spatio-temporelles locales, principe observé chez la mouche, afin de restituer une information de mouvement global relativement insensible au contraste et à la taille de l'objet en mouvement.

Les récents travaux de [Higgins et al.-05] constituent une très bonne référence concernant l'implémentation silicium de ces techniques de perception du mouvement par corrélation spatio-temporelles. En effet les auteurs ont conçu trois circuits implémentant les trois modèles spatio-temporels les plus répandus. Ce sont ceux développés par : Adelson-Bergen (« AB ») [Adelson & Bergen-85], Barlow-Levick (« BL ») [Barlow & Levick-65], et Hassenstein-Reichardt (« HR ») [Reichardt-61]. A la différence des deux autres modèles, le « AB » présente l'avantage d'avoir une réponse indépendante du contraste dès lors qu'il est supérieur à 40%. Cependant le « AB » possède un problème important de courant de fuite lié à la lumière incidente qui augmente les fuites dans tous les drains ou sources des transistors qui deviennent des photodiodes verticales. Ce sont les transistors des multiplieurs qui en sont essentiellement responsables. Les meilleures performances reviennent finalement au « BL », qui fait preuve d'une meilleure sensibilité au mouvement dans sa direction privilégiée.

Les trois circuits sont capables d'extraire la direction du mouvement dès 5% de contraste. En termes d'encombrement des pixels, ils occupent une surface de : 173×173 µm² en CMOS 1.6µm (108×108 λ²) pour le « AB », 142×142 µm² en CMOS 2µm (71×71 λ²) pour le « BL », et enfin 132×132 µm² en CMOS 1.6µm (82.5×82.5 λ²) pour le « HR ».

[Torralba & Hérault-99] ont quant à eux proposé de mettre en œuvre des filtres spatio-temporel pour extraire l'information de mouvement. Ces filtres sont constitués de réseaux résistifs et R-C. Ils informent correctement sur le contenu spatio-temporel de la scène mais requièrent une surface pixel très importante puisque chaque pixel doit contenir des capacités typiquement de 2pF, ce qui les rend difficilement utilisables.

III.2. Des mesures locales vers une information globale

Parmi les capteurs que nous avons évoqués jusqu'ici, certains comme [Tanner & Mead-86], [Andreou et al.-91], [Delbruck-93], ou [Liu-00], fournissent une perception mouvement global dans la scène. Il est obtenu par agrégation d'informations spatio-temporelles mesurées localement.

Le premier système intégré sur silicium implémentant une mesure de mouvement global est celui de [Tanner & Mead-84a] [Tanner & Mead-84b] destiné aux souris optiques. Ce circuit est basé sur la corrélation horizontale ou verticale de pixels voisins dont les luminances sont supérieures ou inférieures à un certain seuil. Un motif comportant des fronts de contrastes saillants doit être projeté sur le capteur afin qu'il puisse extraire les déplacements locaux pour en fournir une information globale. La performance atteinte pour un rapport optique de 1 est une résolution de 100 pixels/pouce pour une vitesse maximum de 2.0 m/s. Ici aussi chaque pixel fonctionne en mode intégration et des traitements à partir de données en courant sont mis en œuvre. La consommation excessive dû à une source lumineuse continue constitue l'inconvénient majeur de ce circuit.

Dans ce même esprit de mise en œuvre de motifs avec contrastes saillants et spécifiques, [Arreguit et al.-96] ont eux aussi proposé un système pour souris optique. Celui-ci comporte 75 détecteurs de mouvement qui observent la fameuse boule de souris de couleur claire et marquée de points foncés

répartis de façon aléatoire. Le capteur compte entre deux instants suffisamment courts pour s'affranchir du problème de moirage (l'aliasing) le nombre de détecteurs ayant perçu un déplacement élémentaire d'un contraste vers la droite, vers la gauche, et le nombre total de contrastes perçus. La mesure du mouvement est obtenue par le rapport de la différence des nombres de détections vers la droite et vers la gauche, par le nombre total de contrastes perçus. Les pixels fonctionnent essentiellement en mode courant, les contrastes sont détectés par comparaison des photocourants. La taille d'un pixel est de l'ordre de 450×450µm en précision de fabrication 2µm, c'est-à-dire 225×225 λ^2. La résolution du mouvement mesuré est de 315 pixels/cm, pour des vitesses comprises entre 0 et 4.6 cm/s.

III.3. Synthèse des détecteurs

Nous reportons dans le Tableau II.1 suivant la synthèse des différents capteurs du mouvement que nous venons de décrire.

Dans l'optique de l'intégration de ces dispositifs dans un imageur - dont le fonctionnement est régi par une lecture synchrone des pixels - nous avons choisi de préciser certaines caractéristiques :

- leur encombrement (taille réelle et taille relative à la technologie employée),
- le type d'information fournie par le capteur (synchrone / asynchrone, mouvement local ou global, 1D ou 2D, vitesse/mouvement explicite ou pas).

Ces données sont des paramètres de la faculté d'un capteur à être intégré au sein d'un imageur, y compris le coût supplémentaire à prévoir en terme de surface.

Par exemple le pixel « FS » de [Kramer-96], dédié à la perception du mouvement, est environ quarante fois plus gros qu'un pixel classique d'un imageur qui lui occupe une surface de l'ordre de 3×3 µm² en technologie CMOS 0.18µm, c'est-à-dire 16.6×16.6 λ^2. La mise œuvre de pixels spécifiques doit par conséquent se justifier par une forte valeur ajoutée.

Auteurs	*Local/global 1D/2D*	*Sortie*	*Vitesse explicite*	*Taille pixel (techno)*	*Taille pixel relative (λ²)*
Mise en correspondance					
[Horiuchi et al.-91]	local / 1D	async.	oui	-	-
[Kramer-96] : FS	local / 1D	async.	oui	220×220 (2 µm)	110×110
[Kramer-96] : FT	local / 1D	async.	oui	-	-
[Kramer-96] : FTI	local / 1D	async.	oui	166×166 (2 µm)	83×83
[Etienne-Cummings et al-93 et 01]	local / 2D	async.	oui	180×310 (2 µm)	90×155
[Yamada & Soga-03]	local / 1D	async.	oui	570×135 (1.5 µm)	380×90
[Miller & Barrows-99]	global / 1D	async.	oui	-	-
[Higgins et al. -99] : ITI	local / 1D	async.	non	110×120 (1.2 µm)	92×100
[Higgins et al. -99] : FTC	local / 1D	async.	non	128×119 (1.2 µm)	107×100
[Ruedi-96]	local / 1D	async.	non	223×215 (2 µm)	111×107
[Moini et al.-97]	local / 1D	sync.	oui	158×158 (1.2 µm)	~130×130
[Tanner & Mead-84a et 84b]	global / 2D	sync.	oui	-	-
[Arreguit et al.-96]	global / 2D	sync.	oui	~450×450 (2 µm)	~225×225
[Ruedi et al.-03]	local / 2D	async.	oui	69×69 (0.5 µm)	138×138
[Navarro-03]	local / 2D	sync.	oui	40×50 (0.35 µm)	114×142
Différentielle					
[Tanner & Mead-86]	global / 2D	async.	non	~450×450 (1.5 µm)	~300×300
[Deutschmann & Koch-98a]	local / 1D	async.	non	147×270 (2 µm)	73.5×135
[Deutschmann & Koch-98b]	local / 2D	async.	non	112×112 (1.2 µm)	93×93
[Stocker-06]	local / 2D	async.	non	124×124 (0.8 µm)	155×155
Corrélation					
[Benson & Delbruck-91]	local / 1D	async.	non	-	-
[Andreou et al.-91]	global / 1D	async.	non	-	-
[Delbruck-93]	pseudo-local / 3×1D	async.	non	225×225 (2 µm)	~112×112
[Harrison & Koch-00]	global / 1D	async.	non	61×199 (1.2 µm)	50×165
[Liu-00]	global / 1D	async.	non	-	-
[Higgins et al.-05] : AB	local / 1D	async.	non	173×173 (1.6µm)	108×108
[Higgins et al.-05] : BL	local / 1D	async.	non	142×142 (2µm)	71×71
[Higgins et al.-05] : HR	local / 1D	async.	non	132×132 (1.6µm)	82.5×82.5
Pixel « image »	Aucun mouvement			3×3 (0.18 µm)	16×16

Tableau II.1. Synthèse de l'état de l'art des principaux capteurs du mouvement réalisés.

CONCLUSION

Le challenge à relever est de proposer une architecture de capteur qui soit optimisée pour adresser la double fonctionnalité image et mouvement global engendré par le déplacement du dispositif d'acquisition. Cet état de l'art sur la perception du mouvement nous permet de tirer certaines conclusions quant à l'estimation du mouvement global à mettre en place.

Tout d'abord, qu'il s'agisse de systèmes biologiques ou artificiels, le mouvement ne peut être perçu que par la présence de zones contrastées, ou texturées, dans la scène. Nous avons également noté que, localement, l'hypothèse d'un mouvement de translation uniforme permet d'obtenir un très bon compromis entre charge de calcul et pertinence du mouvement décrit.

De plus, dès l'instant où l'on n'a aucun a priori sur le type de scène observée et que l'on doit s'attacher à décrire le mouvement entre deux images consécutives d'une séquence vidéo, comme c'est notre cas ici, les techniques de mise en correspondance de plusieurs pixels paraissent les mieux adaptées. Elles procurent de bons résultats et s'avèrent relativement simples à implémenter. Aussi, nous avons vu que certaines techniques de transformées d'images permettent d'améliorer sensiblement la robustesse de cette mise en correspondance.

Du point de vue de l'intégration silicium au niveau pixel de toute ou d'une partie de la mesure du mouvement, c'est aussi l'approche de mise en correspondance qui nous paraît la plus appropriée car la moins sensible au bruit et aux perturbations lumineuses. La détection dans le plan focal d'une zone contrastée de la scène pour la reconnaître dans l'image suivante nous paraît être l'approche la plus robuste et sûre.

Par ailleurs, nous avons vu que la détection et la mesure du mouvement étaient grandement facilitées par un mode de traitement en temps continu. Ces pixels dédiés et spécifiques peuvent fournir directement une vitesse moyenne de déplacement entre deux points dans une direction donnée. Cependant, ces informations « mouvement » sont asynchrones par nature. Aussi, dans le cas du choix de cette solution de mesure des mouvements pour stabiliser une vidéo, il faudra synchroniser ces informations par rapport à l'acquisition d'images.

Enfin, notre système est un dispositif hybride, entre imageur et circuit de vision, qui doit restituer à la fois l'image et l'information de mouvement inter images. Ces deux fonctionnalités apparaissent fortement antagonistes d'un point de vue intégration silicium au niveau pixel. La première requiert des pixels en mode intégration à trois ou quatre transistors pour une surface pixel minimum, et l'autre nécessite des architectures de pixels et éventuellement un mode de fonctionnement modifié ce qui implique des tailles de pixels supérieures et des moyens accrus pour assurer le développement et la validation du capteur. Cette dernière approche n'est pas compatible avec la rentabilité industrielle.

Or, nous avons remarqué que le mouvement global peut être extrait à partir d'un support d'estimation limité à certaines zones de l'image. Dans ce cas la surface ajoutée à l'imageur initial peut être modérée, tout en bénéficiant des avantages d'un conditionnement de l'information lumineuse au sein du pixel directement dédié à la perception du mouvement. Aussi nous présentons dans le chapitre la technique d'estimation du mouvement global que nous avons développée, basée sur la mesure de mouvements en périphérie des images.

Chapitre III.

STABILISATION VIDEO PAR MESURES LOCALES PERIPHERIQUES

INTRODUCTION

Nous décrivons dans ce chapitre la technique que nous avons mise en place pour estimer le mouvement global inter trames d'une séquence d'images. La raison d'être de ce chapitre réside dans la validation et la caractérisation de la technique que nous proposons, ceci par une approche de post-traitement d'image, c'est à dire en faisant abstraction de l'architecture matérielle qui accueillera le traitement.

Notre technique consiste à estimer le modèle de mouvement global choisi, à quatre paramètres, à partir des mouvements locaux mesurés sur la périphérie des images.

Notre démarche a tout d'abord consisté à préciser les spécifications du système du point de vue de l'estimation du mouvement, un paramètre déterminant étant l'amplitude des mouvements à considérer.

Nous introduisons ensuite la solution proposée d'un point de vue qualitatif dans un premier temps. Puis nous mettons en équation le système et validons la technique de façon théorique. Nous étudions notamment le comportement du processus d'optimisation moindres carrés en présence d'éléments perturbateurs tels que du bruit de mesure ou des mouvements parasites.

Enfin après avoir établi une procédure de validation mettant en oeuvre des séquences vidéo synthétiques[60] et réelles en adéquation avec les spécifications requises, nous caractérisons notre technique en considérant la précision du mouvement estimé.

I. SPECIFICATIONS DU SYSTEME

L'estimation du mouvement global inter trames que nous devons mettre en place doit satisfaire les contraintes relatives à une utilisation grand public d'un dispositif portable de type téléphone mobile. Ainsi les environnements d'utilisation du capteur seront variés, de même par conséquent que les vidéos de tests à mettre en œuvre pour la validation de notre technique d'estimation du mouvement global.

De plus STMicroelectronics prévoit d'associer un zoom optique progressif dans les imageurs de prochaines générations, ainsi notre modèle du mouvement global devra intégrer ce paramètre afin de prévoir cette évolution.

Une étape importante pour préciser les spécifications de notre système consiste en un premier lieu à caractériser les mouvements globaux inter trames. Nous nous intéressons ensuite à l'opération de fenêtrage afin de déterminer le nombre de pixels que notre capteur d'image devra comporter pour restituer des vidéos de format donné[61].

---

[60] C'est-à-dire des vidéos artificielles caractérisées notamment par un mouvement global inter images connu.

[61] Rappelons en effet que la stabilisation électronique impose que la vidéo produite en sortie du capteur soit de résolution inférieure aux images effectivement acquises par le capteur.

I.1. Analyse des mouvements globaux inter trames

Pour étudier ces mouvements, nous avons choisi de nous baser sur des séquences vidéo « réelles » (Cf. paragraphe III.1. p. 103 pour les conditions de prise de vue et les caractéristiques des séquences). Il s'agit de séquences panoramiques pour lesquelles nous avons déterminé le mouvement global à l'aide de l'algorithme développé par [Odobez & Bouthemy-95] à l'IRISA de Rennes[62].

Cet algorithme met en oeuvre les statistiques robustes, plus précisément les M-estimateurs, dans un processus de résolution incrémental et multi-échelle. Il est utilisé dans divers cadres applicatifs comme le positionnement dynamique de robots sous-marins pour l'inspection et la fouille sous-marine [Spindler & Bouthemy-98], ou l'augmentation de la résolution des images par traitement (« super résolution » [Dekeyser et al.-00]). Il s'agit d'une technique d'estimation de mouvements globaux performante.

Les paramètres de mouvements obtenus pour la séquence « Guzet[63] » sont reportés sur la courbe intitulée « TX » de la Figure III.1 (page 87), et « TY » de la Figure III.2 (page 87). Celles-ci représentent les évolutions des paramètres de translation « TX » et « TY » du modèle de mouvement en fonction du numéro d'image de la séquence vidéo.

On s'aperçoit que dans une vidéo acquise dans des conditions normales, cadencée à 15 images/seconde, l'amplitude des mouvements globaux entre deux images consécutives est de l'ordre de 1 à 2% de la taille image en moyenne, et peut atteindre 5% au maximum. Dans des conditions de prise de vue particulières, comme celle d'une caméra fixée sur un scooter roulant sur route accidentée par exemple, cette amplitude peut atteindre une valeur moyenne de 2 à 3%, avec des pics jusqu'à 10% de la taille image. Nous ne considèrerons pas ces cas particuliers, et nous attacherons dans un premier temps à stabiliser des vidéos moins chahutées, pour lesquelles l'amplitude de la composante « parasite » des mouvements globaux inter images sera limitée à 5% de la taille image pour une cadence d'acquisition de 15 im/s.

Cependant, la cadence vidéo que nous visons est de 25 à 30 im/s[64], de manière à être capable d'adresser les deux normes PAL/SECAM et NTSC. L'amplitude des mouvements globaux inter images à compenser est alors réduite proportionnellement à l'augmentation de vitesse. Nous considérons donc la stabilisation de déplacements ne dépassant pas 3% de la taille image.

D'un autre côté, nous avons à estimer le plus petit mouvement indésiré « acceptable » après stabilisation. Pour cela, nous avons créé artificiellement des séquences vidéo en fixant le mouvement global inter images. Celui-ci comportait un mouvement global moyen plus un mouvement global indésiré.

---

[62] Le code source C++ de cet algorithme ainsi que les ''makefiles'' associés sont disponibles sur le site internet de l'IRISA de Rennes : http://www.irisa.fr/Vista/Motion2D/index.html.

[63] La séquence dite « Guzet » est une séquence panoramique typique d'une utilisation grand public. Nous en donnons quelques extraits en pages annexes.

[64] La norme européenne PAL/SECAM impose une cadence de 25 images/seconde, et la norme américaine NTSC de 30 images/seconde.

Finalement, nous sommes arrivés à un résultat qui a pour lui la force de l'évidence, même s'il était nécessaire de le vérifier : pour rester imperceptible, le mouvement indésiré doit avoir une amplitude inférieure au pixel. Soit, dans le cas de vidéos au format 320×240, une amplitude inférieure à 0.3% de la taille image.

Finalement, nous devons donc élaborer un système qui, à partir des composantes de translation du mouvement global inter images, devra segmenter le mouvement en deux composantes. La première sera appelée le mouvement « intentionnel » et la seconde sera le mouvement « indésiré ». Cette dernière aura, par hypothèse, une amplitude inférieure à 3% de la taille de l'image. Enfin, la stabilisation sera effectuée en recadrant l'image. C'est la composante « mouvement indésiré » qui fixera la position du centre du cadre par rapport au centre de l'image « brute ». Puisque en deçà d'une amplitude d'un pixel le mouvement indésiré est imperceptible, la position du centre du cadre pourra être arrondie au pixel le plus proche.

I.2. Moyenne temporelle et recadrage

Notre perception visuelle est telle que, pour qu'une séquence vidéo nous paraisse fluide et agréable, elle doit posséder un mouvement inter images sans variations brusques. Aussi nous avons choisi de déterminer le mouvement global intentionnel, ou moyen, par une opération de filtrage temporel passe-bas du mouvement effectif.

Le filtre choisi est un « filtre moyenneur » (ou « filtre exponentiel ») d'équation aux différences :

Eq. III.1. $$\vec{y}[n] = \alpha.\vec{x}[n] + (1 - \alpha).\vec{y}[n - 1]$$

où y[n] et y[n-1] sont les sorties du filtre pour les échantillons « n » et « n-1 ». α est la constante réelle qui représente le poids du filtre ; c'est-à-dire l'influence, sur la sortie y[n], de la nouvelle mesure x[n]. Plus la valeur de α est grande, plus cette nouvelle mesure est prépondérante. Il s'agit d'un filtre à réponse impulsionnelle infinie (IIR), que l'on applique à la série de mouvements globaux entre images consécutives.

Soit $X[i]$ la position d'un point d'intérêt de la scène dans l'image i ;

soit $TX[i] = X[i] - X[i-1]$ le mouvement horizontal de ce point d'intérêt lorsque l'on passe de l'image *i*-1 à l'image *i* ;

enfin, appelons $DX[i]$ le mouvement "voulu", qui est la valeur de $TX[i]$ estimable à partir des valeurs précédentes de *TX*.

Le mouvement parasite, à corriger, est alors : $CX[i] = TX[i] - DX[i]$

avec : $DX[i] = \alpha \cdot TX[i] + (1-\alpha) \cdot DX[i-1]$

On pose alors simplement (Eq. III.2.) : $X_{STAB}[i] = X[i] - CX[i]$

qui est l'image recalée pour stabilisation.

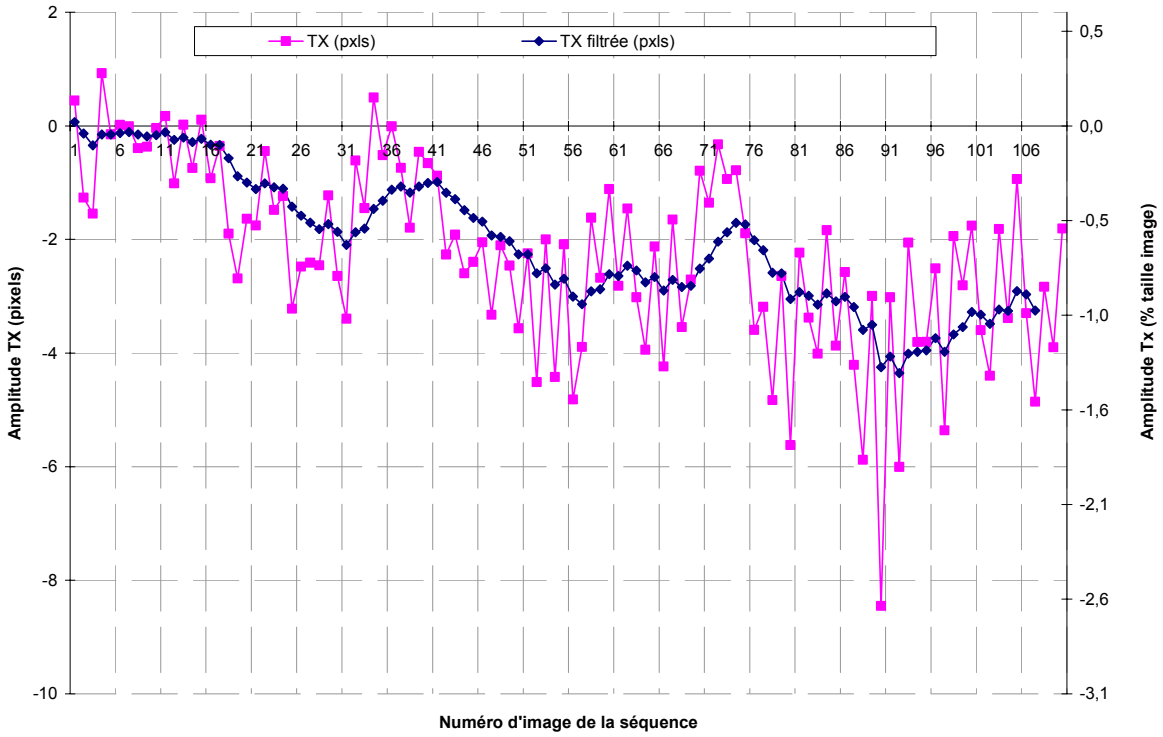


Figure III.1. Évolution de l'amplitude des translations horizontales TX des mouvements globaux inter images pour la séquence « Guzet ».

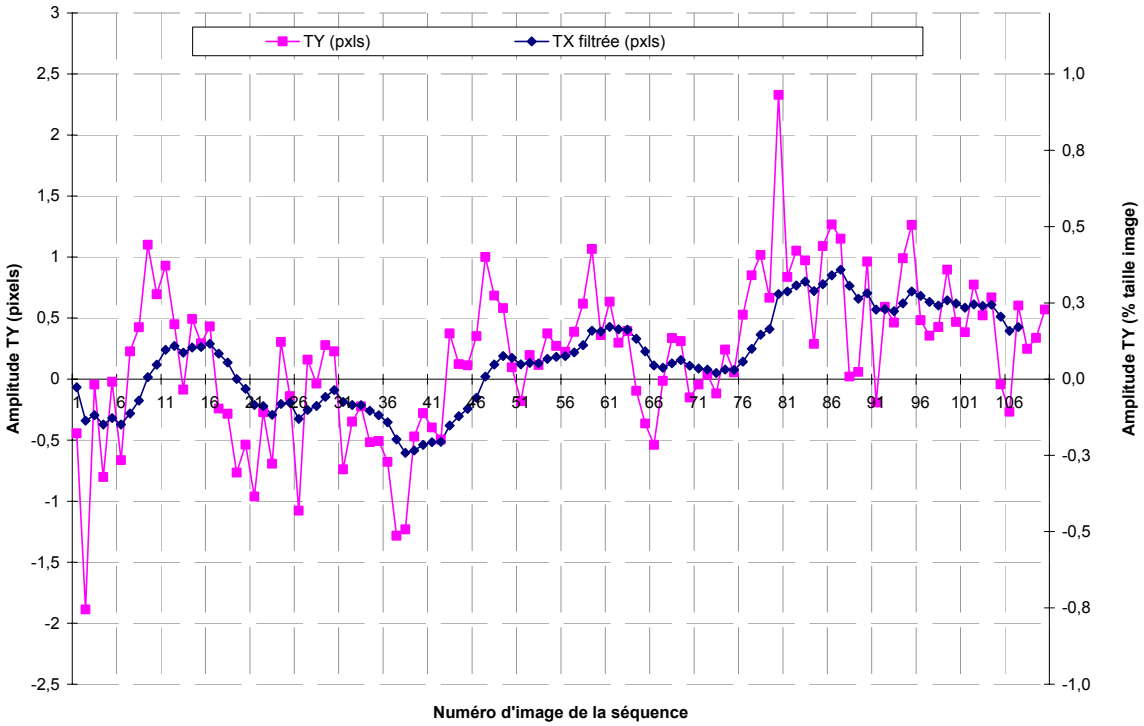


Figure III.2. Évolution de l'amplitude des translations verticales TY des mouvements globaux inter images pour la séquence « Guzet ».

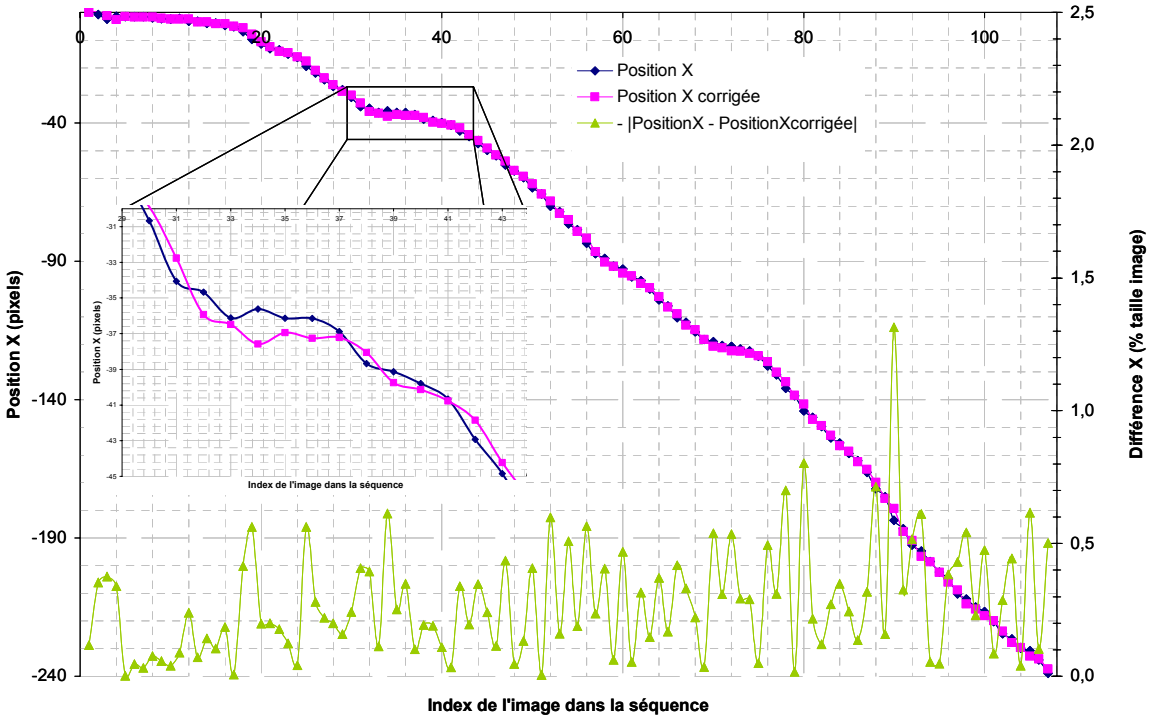


Figure III.3. Évolution des coordonnées horizontales des centres images de la vidéo acquise (PositionX) et stabilisée (PositionXcorrigée), ainsi que leur différence absolue pour la séquence « Guzet ».

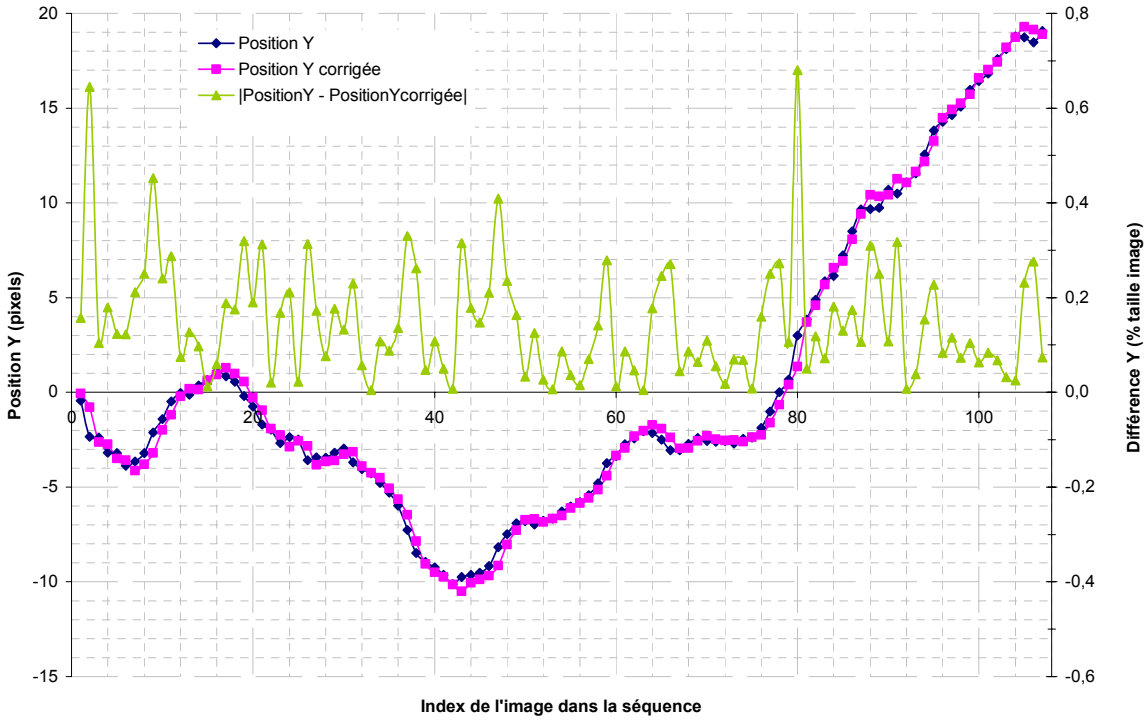


Figure III.4. Évolution des coordonnées verticales des centres images de la vidéo acquise (PositionY) et stabilisée (PositionYcorrigée), ainsi que leur différence absolue pour la séquence « Guzet ».

Nous avons donc appliqué ce filtre sur plusieurs séquences de mouvements globaux inter images extraits de vidéos réelles. On a reporté sur la Figure III.1 et sur la Figure III.2 l'évolution de l'amplitude des mouvements globaux de translations inter images pour la séquence « Guzet »[65].

La courbe intitulée « TX » (resp. TY) représente l'évolution du mouvement de translation horizontal (resp. vertical) mesuré, à l'aide de l'algorithme de [Odobez & Bouthemy-95], entre deux images successives, tandis que la courbe intitulée « TX filtrée » (resp. TY filtrée) représente l'évolution du mouvement de translation « intentionnel » *DX* (resp. *DY*) estimé par notre filtre moyenneur (α = 0,15).

Nous avons validé la valeur du coefficient α en réalisant une série d'essais de filtrages sur différentes séquences de mouvements globaux pour des valeurs de α comprises entre 0.05 et 0.7. Pour chaque essai, nous avons calculé les quantités $CX[i]=TX[i]-DX[i]$ et $CY[i]=TY[i]-DY[i]$ qui sont les composantes du vecteur mouvement global « indésiré » utilisé pour effectuer le recadrage de l'image. Après avoir visualisé les différentes vidéos résultantes, nous avons décidé qu'une valeur de poids du filtre de 0,15 procure les meilleurs résultats. Il est cependant important de noter que cette valeur est qualitative et pourra être modifiée en fonction des critères de performance choisis.

On a représenté sur la Figure III.3 et sur la Figure III.4 l'évolution de la position horizontale (resp. verticale) dans l'image d'un point d'intérêt de la scène (la coordonnée 0,0 correspondant à la position de ce point au début de la séquence) avant et après correction (Cf. Eq. III.3). Ces courbes illustrent bien l'absence d'erreur cumulative. Cependant, l'échelle ne permet pas d'estimer aisément l'amplitude des erreurs instantanées (le mouvement « indésiré »). On a donc reporté sur le même graphique, en dilatant l'échelle, la valeur absolue de la différence entre la position du point d'intérêt dans l'image initiale et dans l'image recadrée. La valeur maximale de cette différence nous donne la largeur d'image additionnelle à prévoir sur les côtés correspondants (gauche ou droite pour X, haut et bas pour Y) pour pouvoir réaliser la stabilisation.

On remarque que dans le cas d'un mouvement panoramique horizontal, il s'agit du paramètre du mouvement décrivant les translations horizontales qui engendre la plus grande différence. Celle-ci atteint un maximum en valeur absolue égal à 1.3% de la taille image horizontale de sortie. Quant au paramètre décrivant les translations verticales, il ne dépasse pas 0.6% de la taille image verticale de sortie.

Nous sommes donc bien dans le cas où l'amplitude des mouvements indésirés reste inférieure à 3% de la taille image.

Appelons « h » et « v » les tailles horizontale et verticale de la vidéo de sortie. Soit « χ » la proportion d'image ajoutée sur chacun des côtés. La surface de l'imageur doit alors être :

Eq. III.3. $$S=(1+2\chi)\cdot h\cdot(1+2\chi)\cdot v=(1+2\chi)^2\cdot S_{vidéo}$$

En posant $\chi=3\%$ on obtient $S=(1{,}06)^2\cdot S_{vidéo}\approx 1{,}12\cdot S_{vidéo}$

Nous y reviendrons au chapitre IV, mais nous pouvons d'ores et déjà constater qu'une correction de mouvements parasites dont l'amplitude atteint 3% de la taille de l'image requiert une augmentation de 12% (quatre fois plus) du nombre de pixels, c'est à dire de la surface de la matrice imageur.

[65] *Cette séquence de 106 images à cadence vidéo de 15 im/s dure 7 secondes environ.*

Cependant, l'approche de stabilisation électronique ne nécessite pas de mécanismes ou de matériel optique particuliers, ce qui implique un encombrement réduit, une meilleure intégration sur la puce, et par conséquent un coût minimum malgré le surcoût en surface silicium lié au recadrage. Elle constitue par conséquent une solution mieux adaptée aux applications grand public.

II. TECHNIQUE D'ESTIMATION DU MOUVEMENT GLOBAL PROPOSEE

Les enseignements que nous avons tirés de l'étude des mécanismes vivants et artificiels de perception du mouvement nous amènent à penser une solution à l'estimation du mouvement global en adoptant une approche qui associe les techniques mathématiques d'estimation ainsi que les architectures spécifiques des capteurs et systèmes photosensibles.

II.1. Principe

Les séquences vidéo, et plus généralement les images, comportent bien souvent trop de détails pour en extraire facilement une information. De nombreux acteurs scientifiques en vision active préconisent en ce sens de favoriser certaines zones ou régions de l'image [Mitiche & Bouthemy-96] [Promising_directions_in_active_vision-91].

Nous avons souligné au chapitre précédent que le mouvement global est directement perçu sur l'arrière plan fixe de la scène et qu'il ne peut l'être aussi aisément sur des objets mobiles dans la scène. Ainsi, la manière la plus évidente d'estimer ce mouvement global est de mesurer le mouvement de l'arrière plan fixe de la scène.

Le cadre de notre application est celui de l'acquisition vidéo « grand-public ». Nous nous sommes donc intéressés à diverses séquences d'images acquises dans ce contexte. Nous avons remarqué que, dans le cas général, la majorité des pixels situés sur les bords des images de ces séquences contiennent l'arrière plan de la scène. En effet un objet mobile dans la scène, c'est-à-dire animé d'un mouvement « parasite » pour nos préoccupations, ne perturbe qu'un nombre limité de pixels. Une grande proportion de l'ensemble des pixels périphériques porte alors l'information utile. La Figure III.5 suivante illustre ce constat et met en évidence notre zone d'intérêt périphérique.

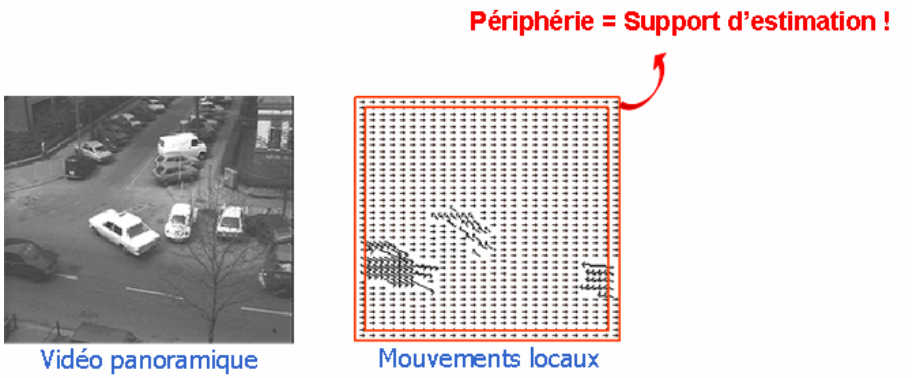


Figure III.5. Image extraite d'une vidéo de type panoramique horizontal, avec son champ de vecteurs mouvements locaux. On remarque que la périphérie de l'image contient majoritairement l'information du mouvement global horizontal.

La solution que nous développons consiste donc à extraire le mouvement global à partir de l'ensemble des déplacements des pixels de cette zone d'intérêt périphérique. Notre support d'estimation du mouvement global sera alors la périphérie des images.

Notons que certains auteurs comme [Sandini et al.-00] et [Etienne-Cummings et al.-00] soulignent cette présence de l'arrière plan sur la périphérie des images et l'utilisent pour réaliser du suivi d'objets. Cette démarche qui consiste à considérer préférentiellement la périphérie des images a également été mise à profit par plusieurs auteurs. [Zahnd et al.-03] s'intéressent à cette zone pour détecter et compter le nombre d'objets entrant et sortant (typiquement des personnes) de la scène visualisée. [Vella et al.-02] utilisent eux aussi les bords et le centre des images pour réaliser de la stabilisation vidéo par post-traitement des images acquises.

En considérant la périphérie des images, la complexité des algorithmes mis en jeu évolue de manière linéaire avec la taille image. Ainsi à même charge de calcul, il sera possible de mettre en œuvre des algorithmes plus robustes que dans le cas d'une solution se basant sur la totalité des pixels qui implique une complexité quadratique et donc des algorithmes plus simples. Cependant ceci ne sera un argument que si notre support d'estimation est pertinent et nous permet d'obtenir une estimation du mouvement global suffisamment précise pour notre application de stabilisation vidéo.

Le fait d'adopter un support d'estimation périphérique présente un autre avantage important : celui d'une bonne répartition des mesures sur les images. En effet, cette disposition contraint particulièrement bien le processus d'estimation des paramètres du mouvement global, notamment ceux de la rotation et du facteur d'échelle.

Nous avons présenté au chapitre précédent plusieurs solutions pour déterminer ces déplacements locaux et nous avons remarqué que, dans le cadre d'une implémentation en post-traitement des images, on doit considérer au cours du temps un voisinage du point où l'on veut déterminer le mouvement. Ainsi les pixels périphériques dont nous parlons ici se situent légèrement à l'intérieur des images afin de permettre de considérer leur voisinage et ainsi estimer leur déplacement d'une image à une autre.

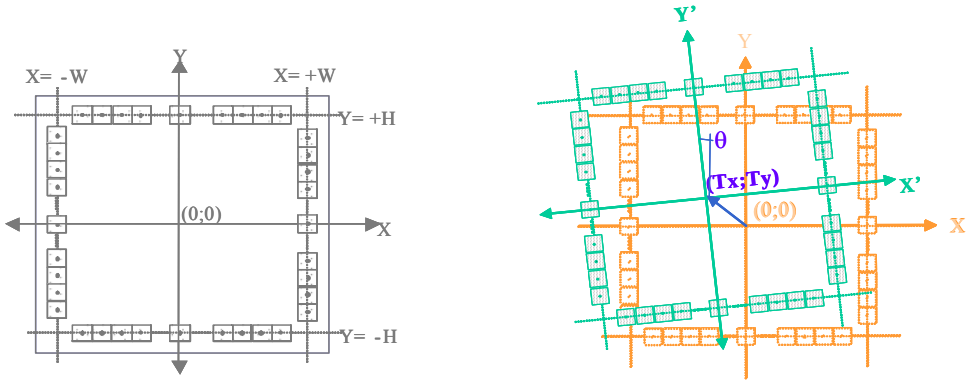
Dans le cadre d'une implémentation matérielle de détecteurs spécifiques de perception du mouvement[66], seuls deux ou trois photodétecteurs voisins sont nécessaires pour fournir les mouvements locaux. Cela autorise par conséquent leur placement sur les bords des images tout en limitant la surface silicium requise et sans pénaliser les pixels dédiés à l'acquisition d'image.

II.2. Formalisation

Le support d'estimation du mouvement global inter images d'une séquence vidéo étant fixé, il nous reste à formaliser le problème du passage d'un ensemble de mouvements locaux mesurés en périphérie de l'image à un mouvement global. Ce n'est qu'après que nous pourrons procéder à sa résolution.

Les données d'entrée de notre système sont les positions des pixels périphériques aux instants « t » et « t+dt », qui sont connues et correspondent respectivement aux images « *n* » et « *n*+1 ».

Nous représentons sur le schéma de la Figure III.6 ci-dessous la transformation géométrique décrivant le mouvement des pixels d'une image animée d'un mouvement de translation et de rotation autour de son centre.



$$\begin{cases} X_j[n+1] = \alpha \cdot \cos\theta \cdot X_j[n] + \alpha \cdot \sin\theta \cdot Y_j[n] + T_X \\ Y_j[n+1] = -\alpha \cdot \sin\theta \cdot X_j[n] + \alpha \cdot \cos\theta \cdot Y_j[n] + T_Y \end{cases}$$

Figure III.6 : Formalisation du problème.

En complément des transformations directement illustrées sur la Figure III.6 que sont la rotation d'un angle θ et les deux translations d'amplitudes *Tx* et *Ty*, nous considérons aussi la transformation de zoom α (ou facteur d'échelle).

Un pixel « *j* » de coordonnées cartésiennes $(X_j[n], Y_j[n])$ dans l'image « n » devient alors le pixel de coordonnées $(X_j[n+1], Y_j[n+1])$ dans l'image « n+1 ». La transformation correspondante est décrite par le système à deux équations de la Figure III.6 dans lequel les paramètres *Tx*, *Ty*, *α* et *θ* sont communs à tous les pixels et dépendent, bien évidemment, de *n*.

[66] *A l'exception du cas de la mise en correspondance spatiale où un voisinage doit là aussi être considéré*

En considérant ces transformations pour chacun des N[67] pixels de la zone d'intérêt et en notant :

$$P = \begin{pmatrix} X_1[n+1] \\ Y_1[n+1] \\ \vdots \\ X_j[n+1] \\ Y_j[n+1] \\ \vdots \\ X_N[n+1] \\ Y_N[n+1] \end{pmatrix},\ K = \begin{pmatrix} 1 & 0 & X_1[n] & Y_1[n] \\ 0 & 1 & Y_1[n] & -X_1[n] \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & X_j[n] & Y_j[n] \\ 0 & 1 & Y_j[n] & -X_j[n] \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & X_N[n] & Y_N[n] \\ 0 & 1 & Y_N[n] & -X_N[n] \end{pmatrix} \text{ et } M = \begin{pmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \end{pmatrix} = \begin{pmatrix} T_X \\ T_Y \\ \alpha\cos\theta \\ \alpha\sin\theta \end{pmatrix}$$

on obtient un système linéaire surdéterminé qui s'écrit simplement :

Eq. III.4. $$P = K \times M$$

où P est le vecteur des coordonnées des pixels périphériques de l'image n+1 dans le repère lié à l'image n. K est la matrice qui code deux fois les coordonnées de chacun des pixels de la zone d'intérêt dans l'image n (les deux premières colonnes indiquent si la ligne sert à calculer une abscisse ou une ordonnée). Enfin, M décrit le mouvement global et représente l'inconnue de notre système.

Remarque : les estimateurs locaux de mouvement ne nous donneront pas directement les $(X_j[n+1], Y_j[n+1])$, mais plutôt un vecteur mouvement qu'il suffira d'ajouter à $(X_j[n], Y_j[n])$ pour obtenir ces nouvelles coordonnées. Ceci aura une influence sur les amplitudes de mouvement à considérer : nous y reviendrons dans la partie consacrée à la caractérisation du processus d'estimation du mouvement global.

De plus, pour simplifier la réalisation physique des estimateurs locaux de mouvement, et surtout pour minimiser leur taille, nous ne prendrons en compte que les mouvements parallèles au bord sur lequel se trouve l'estimateur de mouvement, comme l'illustre la Figure III.7.

Ces mesures unidimensionnelles nous permettent aussi de réduire la complexité du problème de minimisation en divisant par deux la taille des matrices mises en jeu. Pratiquement, chaque nouvelle position dans l'image « n+1 » se caractérise par sa seule nouvelle abscisse ou ordonnée suivant que le bord considéré est horizontal ou vertical.

[67] *Avec N_X points de mesure sur chacun des bords haut et bas et N_Y points de mesure sur chacun des bords droit et gauche, on a $N = 2N_X + 2N_Y$*

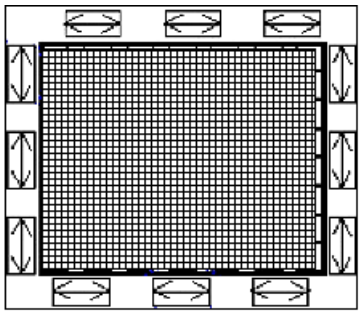


Figure III.7. Mouvements locaux 1D considérés

II.3. Extraction des paramètres globaux : un problème d'optimisation

Le système d'équations à partir duquel nous voulons calculer *M* est surdéterminé, c'est à dire qu'il possède plus d'équations (*N*) que d'inconnues (4). Nous savons que, pour qu'un tel système admette une solution unique, il faut que *N*-4 équations soient linéairement dépendantes des quatre autres. Malheureusement, cela n'est possible qu'en l'absence d'erreurs de mesure sur les mouvements locaux et si la scène observée est indéformable, plane et parallèle au plan image.

La méthode de résolution d'un tel problème a été élaborée par Gauss au 18^{e} siècle lorsqu'il étudiait le mouvement des planètes : puisque toutes les équations du système ne peuvent pas être satisfaites simultanément, on ajoute à chacune une variable qui représente l'écart entre le membre de gauche et le membre de droite de l'égalité originale. Avec *N* équations et *N*+4 variables, le système devient sous déterminé et possède une infinité de solutions. Bien entendu, la solution qui nous intéresse est celle pour laquelle les équations originales sont les plus proches d'être satisfaites. On transforme alors la résolution du système en un problème de minimisation sous contraintes. On peut obtenir une solution analytique pour un système linéaire si l'on choisit comme fonction à minimiser la somme des carrés des *N* variables d'écart qui ont été rajoutées pour le sous déterminer.

La méthode est classique et présente plusieurs avantages :l'estimateur « moindres carrés » est relativement simple à implémenter et présente un coût de calculs réduit, il est non biaisé dans le cas d'un modèle de mesures linéaire Gaussien de moyenne nulle [Kay-93]. Enfin, et ce n'est pas le moins important, sa solution est explicite sous la forme :

$$M = \left(\left(K^T \times K\right)^{-1} \times K^T\right) \times P \qquad \textit{Eq. III.5.}$$

Dans le cas présent, la matrice *K* est invariante d'une image à l'autre, car elle code les positions des points où sont mesurés les mouvements locaux. Il s'ensuit que la matrice $K' = \left(K^T \times K\right)^{-1} \times K^T$, pseudo-inverse de *K*, peut être déterminée une fois pour toutes et le calcul de *M* est immédiat.

Cependant l'estimation doit être robuste aux mouvements singuliers et parasites d'objets mobiles dans la zone d'intérêt. Aussi, nous avons associé à chaque mesure locale du mouvement un poids de confiance en la mesure, nous menant à définir une matrice de confiance *W*, diagonale, qui contient ces poids. La solution correspond alors à une minimisation des erreurs au sens des moindres carrés pondérés. Celle-ci est donnée par :

$$\text{Eq. III.6.} \qquad M = \left(\left(K^T \times W \times K\right)^{-1} \times K^T \times W\right) \times P$$

La matrice *W* étant maintenant variable d'une image à l'autre, nous perdons l'avantage souligné précédemment. Nous montrerons cependant au chapitre IV que la complexité du calcul de *M* reste en O(*N*) (c'est à dire une fonction affine du nombre de points de mesure).

En pratique, la matrice W pourra être déterminée à partir de deux grands types d'informations. D'une part, les estimateurs locaux de mouvement pourront assortir leur mesure d'un indice de confiance dépendant essentiellement des caractéristiques de l'image à l'instant et au point considéré (une zone uniforme ou moyennement texturée ne permet pas d'estimer le mouvement avec la même précision que lorsqu'on est en présence d'un front de contraste bien marqué). D'autre part, les poids constituant la matrice *W* pourront aussi être déterminés *a posteriori* pour éliminer les mesures par trop atypiques. Supposons, par exemple, qu'un détecteur local milite pour une translation à gauche alors que la majorité des autres indique une translation à droite, il y a fort à parier que le détecteur atypique observe en fait un objet mobile de la scène. On obtiendra donc une meilleure estimation du mouvement global en éliminant cette mesure.

Cette détermination *a posteriori* des poids des mesures élémentaires pourra se faire en deux étapes. D'abord, une première estimation du mouvement global sans pondération, ou bien avec une pondération ne prenant en compte que les indices de confiance donnés par les détecteurs permettra de calculer le mouvement que *devrait* percevoir chaque détecteur. En comparant ce mouvement estimé et le mouvement effectivement mesuré, on va pouvoir corriger la matrice *W* de façon à minimiser l'influence des détecteurs fournissant une mesure trop atypique. Il ne reste plus alors qu'à refaire l'estimation du mouvement global au sens des moindres carrés pondérés.

La procédure d'estimation est itérative. En pratique, comme illustré sur la Figure III.8 suivante, nous réalisons une ou deux itérations. Le vecteur M2 résultat de l'estimation contient les données décrivant les paramètres du mouvement estimé (Tx, Ty, α.cos θ, α.sin θ), d'où on déduit θ et par suite α.

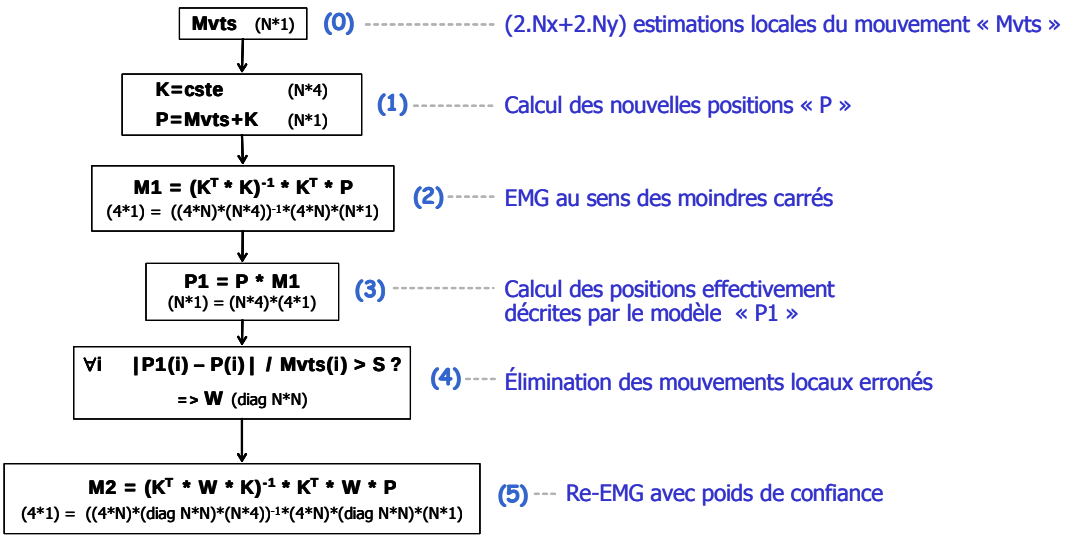


Figure III.8. Procédure d'estimation du mouvement global.
les valeurs entre parenthèses représentent les tailles des matrices mises en jeu dans le calcul.

II.4. Caractérisation théorique

Avant de mettre en place toute la procédure d'estimation du mouvement global, partant de l'image et passant par l'estimation locale des mouvements en périphérie, il nous faut caractériser la méthode que nous proposons pour passer des mouvements locaux périphériques au mouvement global.

Comme l'illustre la Figure III.9, les caractérisations menées ici sont mises en place en appliquant une transformation géométrique aux positions de l'image de référence « *n* » (« + » sur Figure III.9) en accord avec le modèle de mouvement à estimer (*Tx*, *Ty*, α, θ). On obtient alors la nouvelle image « *n*+1 » (« o » sur Figure III.9), ainsi que les vecteurs mouvements locaux idéaux et théoriques dont nous ne considérons que la composante parallèle au bord (mouvement local 1D, ➙ sur Figure III.9). Puis à partir de ces vecteurs mouvements locaux théoriques, nous estimons le modèle du mouvement global au sens des moindres carrés, comme décrit en section II.3. par l'équation Eq. III.5.

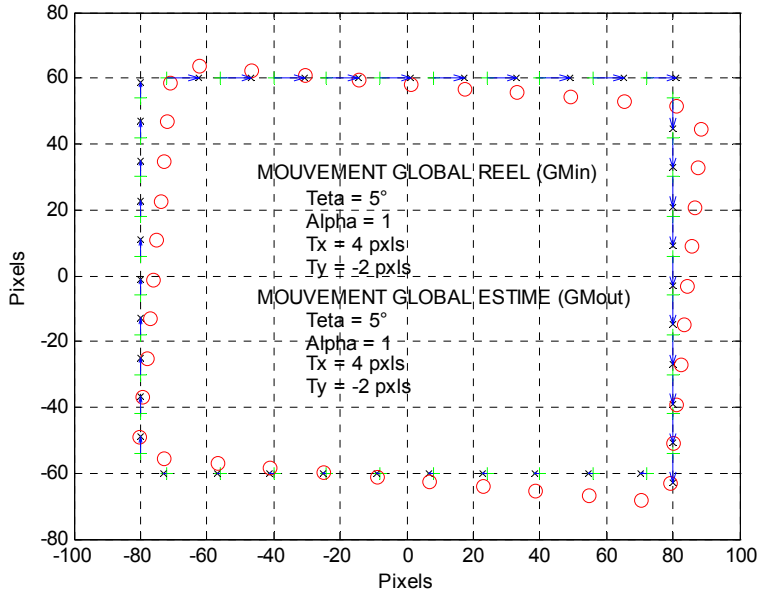


Figure III.9. Exemple de simulation de l'estimation du mouvement global
+ : positions des estimateurs de mouvement locaux dans l'image « n »
o : positions des points suivis par les estimateurs de mouvement locaux dans l'image « n+1 »
x : projections sur les côtés des points suivis dans l'image « n+1 »
➙ : vecteurs mouvements locaux 1d entre les images « n » et « n+1»

II.4.a. Dynamique et linéarité

Les quatre variables que nous estimons à l'aide du processus décrit ci-dessus sont :

$$M_1 = T_X$$
$$M_2 = T_Y$$
$$M_3 = \alpha \cos\theta$$
$$M_4 = \alpha \sin\theta$$

Ainsi les deux premiers paramètres du modèle de mouvement global que sont *Tx* et *Ty* s'obtiennent directement, donc sans limite théorique en dynamique, sinon celle des dispositifs de mesure.

Quant au facteur d'échelle α et à l'angle de rotation θ, ils se déterminent par :

$$\alpha = \sqrt{M_3{}^2 + M_4{}^2} \qquad \text{et} \qquad \theta = \arctan\left(\frac{M_4}{M_3}\right)$$

Le facteur d'échelle ne possède donc pas non plus de limitations en dynamique et linéarité, ce qui n'est pas le cas pour l'angle de rotation qui lui doit être compris entre -90° et +90° pour être estimé sans ambiguïté.

En effet, en se basant sur la Figure III.9, si nous faisons croître l'angle de rotation (mouvement du périmètre dans le sens horaire), on remarque qu'une position initiale sur le bord haut voit son abscisse augmenter dans sa nouvelle position (après rotation) jusqu'à ce que l'angle de rotation atteigne 90°. Pour des angles supérieurs, l'abscisse de la nouvelle position se met à décroître. Cette nouvelle abscisse s'avère par conséquent identique pour, par exemple, des angles de rotation de 85° ou 95°. L'angle de rotation estimé doit donc être limité à +90° ou -90°. Ceci est dû au fait que nous ne considérons que des mouvements locaux 1D. Dans le cas 2D, cette ambiguïté n'existerait pas.

La dynamique maximale théorique des mouvements estimables entre deux images n'est donc pas limitée pour les paramètres de translations et de facteur d'échelle, à la différence des rotations qui doivent, quant à elles, être comprises entre -90° et +90°. Cependant, cette limite est très au delà de celle qui nous sera imposée par les estimateurs locaux de mouvement qui ne pourront mesurer que des déplacements de quelques pixels.

De plus les simulations (Cf. Figure III.9) ont montré qu'en l'absence de perturbations, l'estimation est linéaire sur toute cette dynamique, comme nous l'illustrons sur la Figure III.10 ci-dessous.

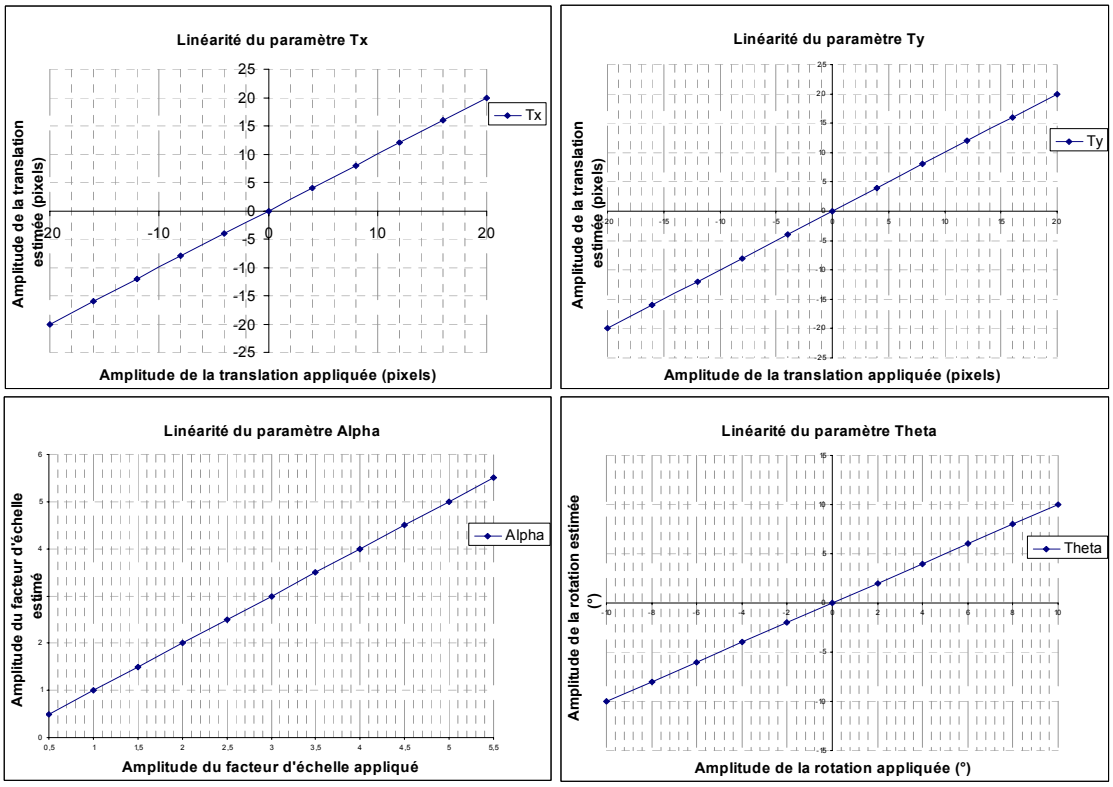


Figure III.10. Étude en simulation de la linéarité de l'estimation des paramètres.

II.4.b. Robustesse au bruit d'estimation des mouvements locaux

Pour caractériser la robustesse de l'optimisation au sens des moindres carrés vis à vis du bruit affectant la mesure des mouvements locaux, nous ajoutons à chaque mouvement local théorique un bruit additif Gaussien. Ce bruit est centré en 0 et d'écart type σ. Puis on étudie, pour une taille d'image donnée et un nombre de mouvements locaux croissant, l'erreur absolue moyenne et l'écart type d'une série de 50 estimations.

On remarque que l'impact du bruit est le même quels que soient l'amplitude et le type des mouvements considérés. Par exemple, avec un bruit additif Gaussien d'écart type de 2% de la taille image, et un nombre croissant de mesures locales du mouvement par côté, on obtient les erreurs d'estimations présentées sur la Figure III.11, indépendamment du mouvement global simulé. On remarque que l'erreur ainsi que l'écart type diminuent avec la croissance du nombre de mouvements locaux considérés pour tendre vers une limite qui n'est autre que l'image de l'incertitude sur les mesures élémentaires.

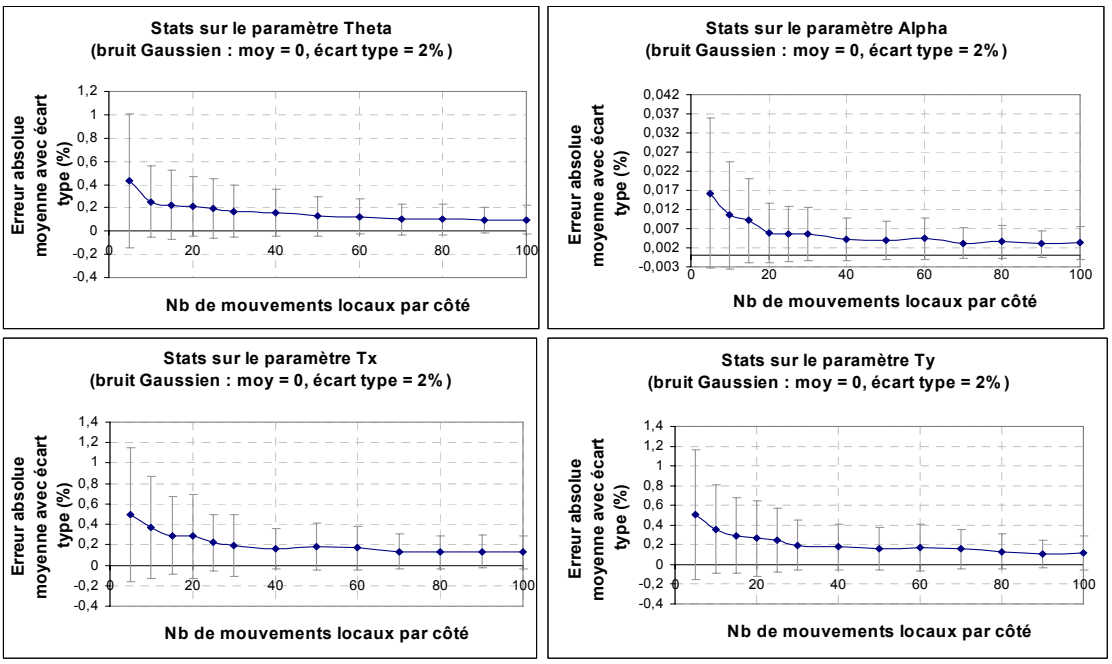


Figure III.11. Erreur absolue moyenne et écart type de l'estimation du mouvement global obtenus sur des séries de 50 estimations, engendrés par un bruit Gaussien (moy=0, écart type=2% de la taille image) ajouté aux mouvements locaux[68].

[68] *Les erreurs absolues moyennes sont toutes exprimées relativement à une grandeur fixe. L'erreur sur la rotation est rapportée à sa dynamique (90°), l'erreur sur le zoom est directement exprimée en pour cent puisqu'il s'agit d'un nombre sans dimension. Enfin, l'erreur sur une translation est rapportée à la taille de l'image.*

II.4.c. Robustesse aux mouvements parasites

On considère ici un mouvement parasite donné (du type objet mobile dans la scène) affectant un nombre donné de mouvements locaux. On observe alors, pour une taille d'image donnée, l'évolution de l'erreur d'estimation pour différentes amplitudes du mouvement parasite et en fonction du nombre de mesures locales par côté.

L'erreur induite par le mouvement parasite est identique quel que soit le mouvement global. L'évolution de cette erreur d'estimation est linéaire en fonction de l'amplitude du mouvement parasite appliqué. Cela signifie par exemple que l'erreur engendrée par un même mouvement parasite sur une image de taille double est deux fois moins importante. De plus, cette erreur décroît asymptotiquement vers zéro en fonction du nombre de mouvements locaux. Nous mettons en évidence ce phénomène sur la Figure III.12.

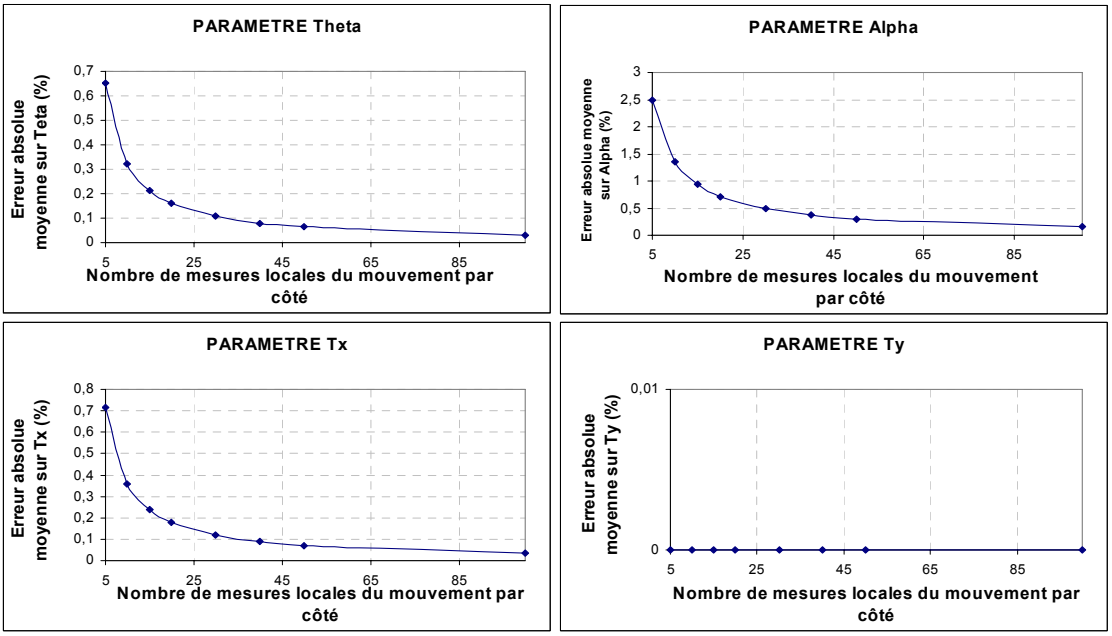


Figure III.12. Evolution de l'erreur absolue moyenne engendrée par un mouvement parasite de type translation en X d'amplitude 7% de la taille image, en fonction du nombre de mouvements locaux.

II.4.d. Conclusion

Les caractérisations décrites dans la présente section ont donc montré qu'il est possible d'estimer le mouvement global d'une image à partir des composantes parallèles aux bords des mouvements locaux périphériques. Cette estimation est linéaire sur une grande dynamique de mouvement, sans erreur systématique. De plus, l'erreur due au bruit de mesure et aux mouvements locaux parasites (produits par des objets mobiles dans la scène) décroît quand on augmente le nombre de points de mesure.

En tenant compte des caractéristiques des estimateurs locaux de mouvement, il sera donc possible d'arriver à un compromis entre le temps de calcul et la précision attendue pour l'estimation du mouvement global.

II.5. Estimations des mouvements locaux périphériques

Notre état de l'art sur les capteurs du mouvement et les différentes techniques d'estimation des mouvements locaux nous a orienté vers les techniques de mise en correspondance qui fournissent un meilleur rapport performance/complexité.

Nous nous intéressons à deux solutions appartenant à cette catégorie pour estimer les mouvements périphériques. La première réalise un appariement de pixel à pixel à partir d'une transformée robuste de l'image, la seconde consiste à extraire et à apparier les contrastes présents dans la zone périphérique de l'image. L'intérêt de ces méthodes est double : elles restent suffisamment simples pour pouvoir être intégrées, au moins partiellement, au niveau du capteur, tout en présentant une robustesse satisfaisante.

Nous incluons également dans notre étude une troisième technique qui consiste à apparier un bloc de pixels d'une image à une autre. Comme nous l'avons souligné, cette technique est reconnue robuste et constituera pour nous une référence de comparaison pour les performances des deux autres techniques.

II.5.a. Appariement de blocs de pixels

La technique d'estimation des mouvements que nous avons utilisé est celle de l'appariement de blocs de pixels, ou « block matching ». Les vecteurs mouvements sont obtenus par mises en correspondance de blocs de pixels. Parmi les nombreuses variantes proposées, [Golston-04] est celle qui donne le résultat optimum. Elle consiste à rechercher, parmi tous les possibles[69] d'une zone donnée de l'image n+1, le meilleur appariement avec un bloc de pixels de l'image n.

La taille des blocs que nous avons choisie est de 5×5 pixels, le critère de minimisation retenu est la somme des différences absolues. Estimer le mouvement d'un bloc dont le pixel central a pour coordonnées (i,j) dans l'image n, consiste alors à rechercher dans l'image n+1 le bloc de pixels minimisant la somme des différences absolues pixel à pixel avec le bloc considéré.

Par exemple, considérons une séquence vidéo caractérisée par un mouvement de translation vers la droite de 3 pixels entre deux images consécutives. Le bloc possédant la plus petite somme des 25 différences absolues pixel à pixel est celui dont le pixel central se trouve en position (i+3,j) dans l'image n+1.

L'appariement de blocs de pixels tel que nous venons de le décrire s'effectue sur l'image originale. Pour rendre l'estimation plus robuste, aux variations de luminosité par exemple, il est possible de « pré-coder » l'image.

II.5.b. Appariement de codes de texture

La première approche que nous avons privilégiée se place dans la continuité de travaux menés au laboratoire [Navarro-03]. Cette technique consiste à apparier un pixel d'une image « I_1 » dans la suivante « I_2 ». Cette mise en correspondance est réalisée à partir d'une transformée spatiale des images qui

[69] *C'est pourquoi cette technique est appelée « Full Search Block Matching, FSBM ».*

attribue à chaque pixel un code en fonction de la texture de son voisinage. Le code d'un pixel ainsi obtenu dans l'image « I_1 » est alors recherché dans l'image « I_2 ».

Cette transformée spatiale, appelé « transformée du recensement »[70] est une transformée locale non paramétrique qui a été initialement proposée par [Zabih & Woodfill-94]. [Banks et al.-97] ont démontré sa robustesse en présence de variations lumineuses importantes qui sont présentes au cours des séquences vidéo, à cause des ombres par exemple. La transformée du recensement est mise à profit dans plusieurs applications où la mise en correspondance robuste de pixel est nécessaire : en stéréo-vision [Woodfill & Hertzen-97], en vidéo-conférence immersive [Schreer et al.-01], en « e-shopping » [Chung et al.-04] et en reconnaissance de visage [Froba & Ernst-04] par exemples. La Figure III.13 suivante illustre ce codage de texture qui consiste à comparer la luminance du pixel considéré, le pixel de luminance 112 sur la figure, par rapport à ses pixels voisins directs. Si la luminance du pixel considéré est supérieure à son voisin, le bit « 0 » est attribué, dans le cas contraire, c'est le bit « 1 ». Le code résultant des comparaisons des huit pixels voisins est donc un octet, c'est-à-dire un espace mémoire semblable à celui d'un pixel de l'image.

Grâce à cette transformée, un bloc de 5×5 pixels est codé par une chaîne de 25 bits qui peut être manipulée très efficacement par un calculateur, accélérant d'autant les appariements.

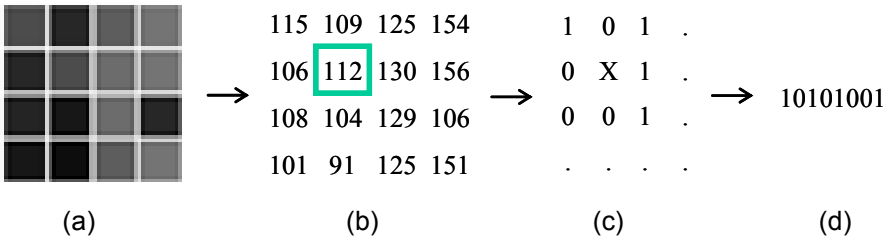


Figure III.13. Codage du recensement.
(a) portion d'image, (b) table des luminances des pixels, (c) tableau recensant, sur un voisinage 3x3, les pixels de luminance plus grand ou plus petite que celle du pixel encadré ; (d) code (8 bits pour un voisinage 3x3) associé, dans la transformée, au pixel encadré.

II.5.c. Appariement de blocs de pixels après extraction de contrastes

Nous venons de voir qu'une façon d'alléger les calculs, et donc d'accélérer l'appariement des blocs de pixels peut être de travailler sur une image seuillée où chaque pixel est codé par un seul bit. Il n'est cependant pas très intéressant de seuiller l'image elle même, en ce sens que ce seuillage peut faire apparaître des artefacts qui fausseront l'appariement des pixels. En revanche, un seuillage du gradient permet de faire apparaître les fronts de contraste de l'image qui sont les points où la mesure du mouvement est la plus fiable.[71]

[70] *« Census transform » en Anglais.*

[71] *Une étude plus précise de la transformée du recensement nous montrerait que le codage de texture qu'elle réalise n'est autre qu'un codage du signe du gradient, ce qui est encore une autre façon de coder les contrastes de l'image.*

Nous avons remarqué lors de notre état de l'art qu'il est possible de réaliser au niveau pixel l'opération de détection de contours. Aussi nous caractérisons ici une technique d'estimation des mouvements locaux s'appliquant sur des séquences vidéo sur lesquelles les contours sont extraits.

Nous réalisons cette détection à l'aide du filtre développé par Canny, qui est reconnu très performant [Montesinos] [Horaud & Monga-95]. La détection de contour résultante sur la séquence « Nature » est illustrée ci-dessous, avec un seuil haut fixé à 0.05 et un seuil bas à 0.02.

Bien évidemment, le filtre de Canny n'est pas simplement intégrable dans le plan focal. Nous verrons au chapitre IV qu'il existe cependant d'autres possibilités, certes moins robustes, pour faire de la détection de contours dès l'acquisition de l'image.

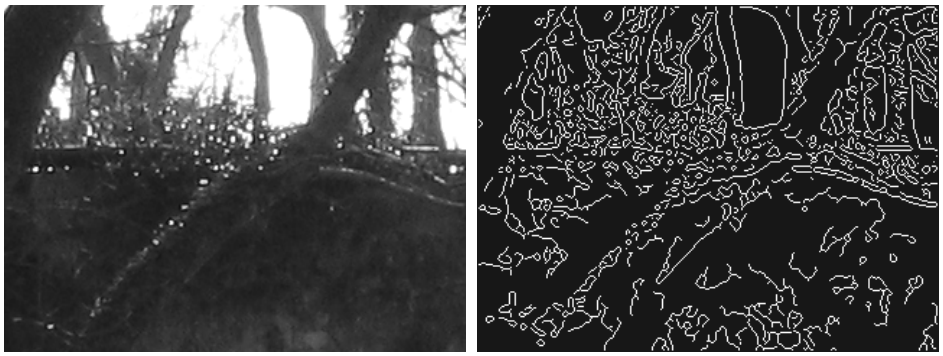


Figure III.14. Détection de contours par filtrage de Canny, séquence « Nature ».

III. PROCEDURE DE VALIDATION

Nous devons caractériser notre technique d'estimation du mouvement dans les conditions les plus proches possibles de celles de sa future utilisation, c'est-à-dire en considérant des séquences vidéo de test qui rendent compte au mieux d'une utilisation grand public du dispositif. Les scènes choisies sont par conséquent variées. Deux types de vidéos tests ont été mis en place pour la caractérisation : des séquences réelles et synthétiques.

De plus, dans la perspective de reporter le traitement ou une partie du traitement au niveau des pixels, dans le plan focal du capteur, nous devons tenir compte du bruit présent à cet étage de la chaîne de traitement du signal. Ainsi nous avons développé un modèle de bruit, que nous avons ajouté aux séquences synthétiques, afin de quantifier son influence sur les performances obtenues.

Le Tableau III-1 récapitule l'ensemble des séquences que nous avons utilisées et leurs caractéristiques respectives.

Séquence	Cadence	Durée	Nb images	Taille	Format
Nature	15 im/s	7 s	106	320×240	MJPEG v3 @ 320kbps
Guzet	15 im/s	7 s	106	320×240	MJPEG v3 @ 320kbps
Bureau	15 im/s	7 s	106	320×240	MJPEG v3 @ 320kbps
Pano_Ext	25 im/s	4 s	100	720×576	DV

Tableau III-1. Séquences d'images utilisées pour nos caractérisations.

III.1. Séquences réelles

Nous avons acquis ces séquences vidéo à l'aide deux caméras. La première est une caméra numérique CANON PowerShot A85. Les vidéos couleurs obtenues sont des fichiers .AVI (Audio Video Interleaved) avec les caractéristiques suivantes :

Format des images	320×280 (CIF)
Format de compression vidéo	Motion-JPEG v3
Débit vidéo	320 kbps (15 images/s)

La seconde caméra numérique que nous avons employée est une Sony, dont les vidéos ont les caractéristiques suivantes :

Format des images	720×576 (PAL)
Format de compression vidéo	DV
Débit vidéo	25 Mbps

L'intérêt d'utiliser deux types de caméras distincts réside dans le fait de tester l'influence de la compression vidéo sur notre algorithme. En effet celle-ci est sévère dans le premier cas, alors qu'elle l'est beaucoup moins dans le second.

Figure III.15 : A gauche : séquence en extérieur, « Nature ».
A droite : séquence en intérieur « Bureaux ».

Les deux séquences réelles que nous considérons ici contiennent d'une part une scène en environnement extérieur très texturé comportant des objets de réflectivité variée, nommée « Nature » (Cf. Figure III.15). Et d'autre part une scène en intérieur avec des motifs plus uniformes comportant là aussi des réflectivités variées, appelée « Bureaux » (Cf. Figure III.15). On peut noter aussi une différence fondamentale au niveau de l'éclairage de la scène. La première est éclairée par la lumière solaire, tandis que la deuxième est éclairée par des tubes luminescents.

Il est important de considérer des vidéos réelles car il existe des effets liés à l'imperfection des systèmes d'acquisition. Le flou de bougé notamment peut nuire aux performances de l'algorithme d'estimation du mouvement.

Afin de quantifier les performances de notre estimation du mouvement, il nous est indispensable de connaître précisément le mouvement global entre deux images consécutives de nos séquences. Nous avons pour cela appliqué l'algorithme de [Odobez & Bouthemy-95]. Les paramètres estimés constituent

alors les mouvements de référence, à partir desquels nous déduirons la précision de nos estimations au paragraphe IV.

III.2. Séquences synthétiques paramétrées

D'autre part, nous avons créé des séquences vidéo artificielles, pour lesquelles nous avons fixé le mouvement inter images. Ces séquences vidéo sont construites en considérant une image référence, haute résolution, à partir de laquelle on extrait des images de taille inférieure, de telle manière que leur succession constitue une séquence vidéo dont le mouvement global inter images est parfaitement connu. Dans ces conditions, les performances des algorithmes d'estimation du mouvement sont quelque peu idéalisées puisque les conditions d'éclairement, les réflexions lumineuses, les bruits liés au dispositif d'acquisition, sont ignorés. Ces séquences présentent l'avantage de pouvoir nous renseigner sur le comportement optimal des algorithmes que nous proposons.

Deux paramètres sont importants ici : l'image référence et le mouvement global appliqué. Les images références choisies doivent rendre compte au mieux des scènes qui seront rencontrées par les utilisateurs. Dans ce cas, les caractérisations de notre technique d'estimation du mouvement seront les plus réalistes possibles. Ensuite, les mouvements globaux inter images que nous avons choisis doivent eux aussi être les plus réalistes possibles. Nous nous sommes alors basés sur l'étude de l'amplitude des mouvements globaux que nous avons menée au paragraphe § I.1.

Les mouvements inter images que nous considérons sont caractérisés par des translations (horizontales, verticales, et diagonales) d'au maximum 3% de la taille image, des rotations autour de l'axe optique de moins de 3°.

IV. PERFORMANCES OBTENUES

A partir des images et séquences vidéo que nous avons définies dans la procédure de validation, nous présentons ici les performances que nous avons obtenues. Dans un premier temps nous caractérisons les estimations des mouvements locaux périphériques, puis celles relatives aux mouvements globaux estimés.

IV.1. Estimation des mouvements locaux périphériques

La métrique que nous avons utilisée ici pour juger des performances de chacune des trois techniques d'estimation des mouvements locaux considère le taux de bons appariements. Nous définissons ce taux comme étant le rapport du le nombre de vecteurs mouvements « justes » par le nombre total d'estimations réalisées :

Eq. .7.
$$Taux_{pixels\ appariés} = \frac{Nombre\_de\_mouvements\_"correctement"\_estimés}{Nombre\_total\_d'estimations}$$

Un mouvement est considéré correctement estimé lorsqu'il est égal, au pixel près, au mouvement effectif.

Nous avons caractérisé les performances de chacune des trois techniques d'appariement en considérant les valeurs du taux de pixels appariés en fonction de l'amplitude de la zone de recherche. En effet, élargir la zone de recherche augmente la dynamique de la mesure de mouvement local, au prix d'un risque plus grand de faux appariement. Cette caractérisation a été faite sur des images de 320×240 pixels.

IV.1.a. Appariement de blocs de pixels

Pour un bloc de pixel, de taille 3×3 ou 5×5, dans une image de la séquence vidéo de test, nous recherchons, dans l'image suivante, le bloc de pixels le moins dissemblant. L'indice de dissemblance de deux blocs de pixels est la somme des valeurs absolues des différence (SAD) de luminance pixel à pixel.

Il s'agit de la technique la plus performante. Elle atteint sans difficulté des scores voisins de 100% sur les séquences synthétiques, quasi indépendamment de la distance ce recherche. La Figure III.16 ci-dessous résume les performances obtenues pour la séquence synthétique « Nature ».

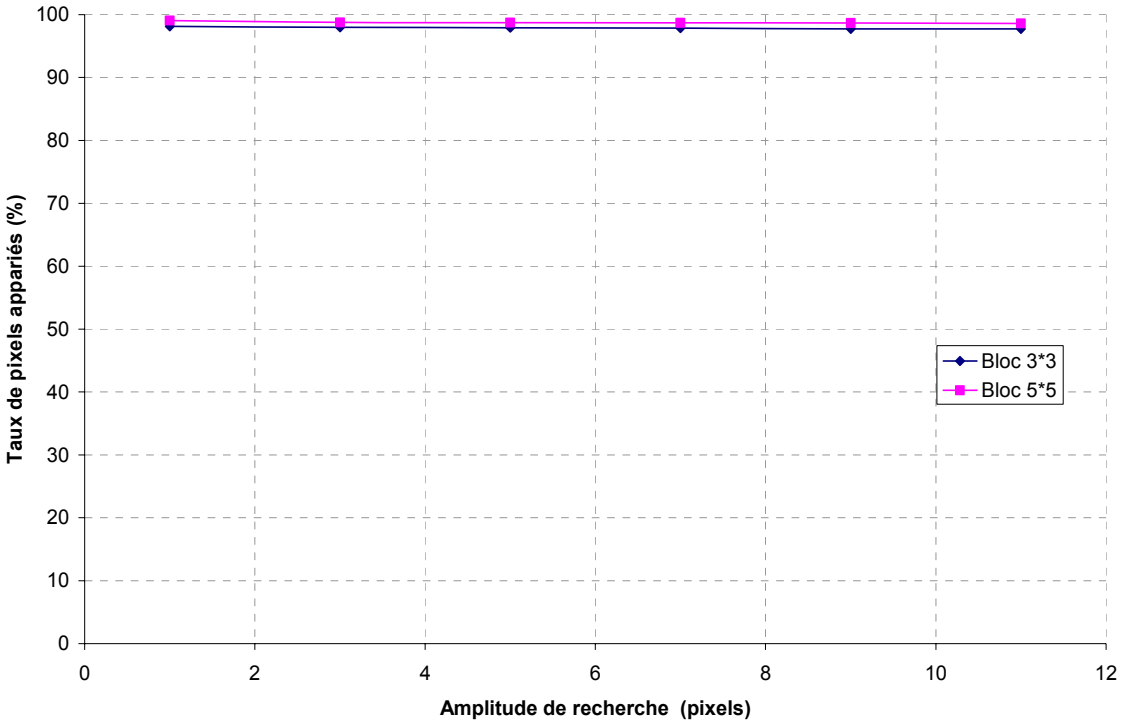


Figure III.16. Taux de pixels appariements en fonction de la taille de la zone de recherche pour une taille de bloc de pixels de 3×3 et 5×5.

IV.1.b. Appariement de codes de texture

Nous réalisons ici les mises en correspondance d'un pixel d'une image à une autre à partir du code de la texture du voisinage du pixel. Deux pixels sont appariés si la distance de Hamming de leurs codes binaires respectifs est minimale. Par conséquent, plus la zone de recherche est grande, plus on a de

chances de rencontrer un code identique au code du pixel à apparier, sans que ce code corresponde réellement au pixel souhaité.

Nous représentons sur la Figure III.17 ci-dessous les performances obtenues sur la séquence vidéo synthétique paramétrée « Nature ».

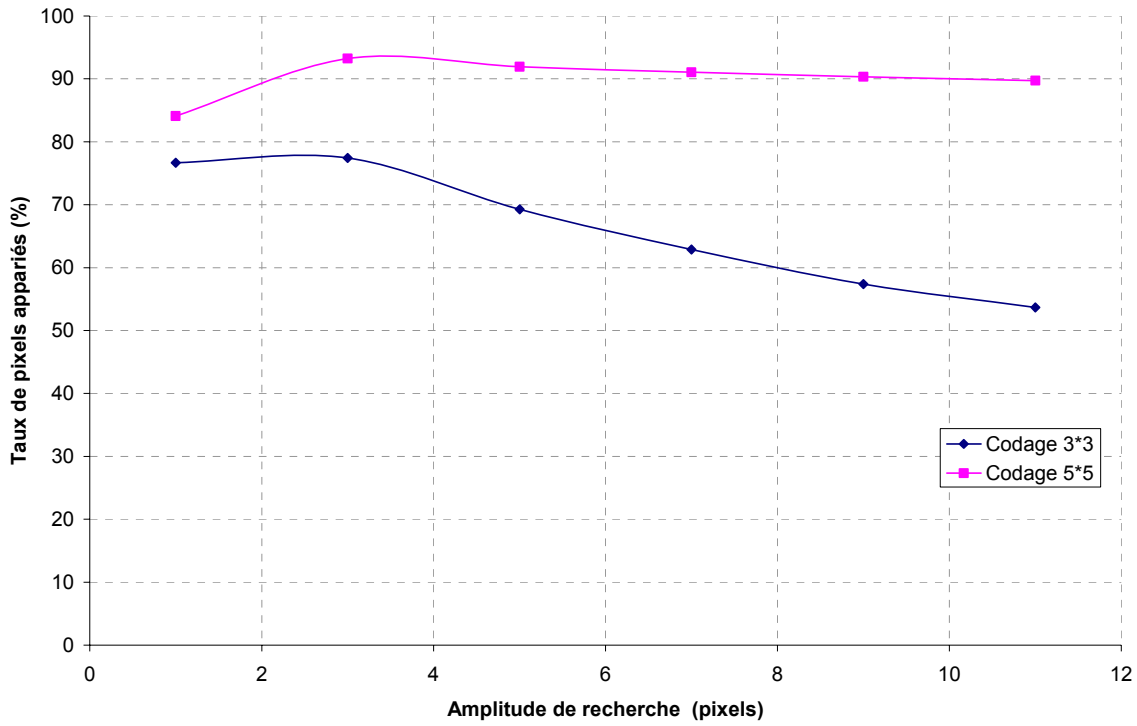


Figure III.17. Taux de pixels appariés en fonction de la taille de la zone de recherche pour un codage du recensement sur un voisinage 3×3 et 5×5.

Nous constatons que le taux de bons appariements décroît avec l'augmentation de l'amplitude de recherche, comme cela était prévisible. Cependant, la taille du voisinage de codage tempère cette décroissance. Ceci est dû au fait que plus le voisinage de codage est large, plus le code binaire de chaque pixel comporte de bits, et moins il y a de chances d'obtenir de faux appariements.

IV.1.c. Appariement de blocs de pixels avec extraction de contrastes

Les images que nous considérons ici sont le résultat d'une extraction de contrastes par le filtre de Canny. On effectue alors un appariement de blocs sur la somme absolue des différences (Cf. IV.1.a.) Nous reportons ci-dessous les performances obtenues pour la séquence synthétique « Nature ».

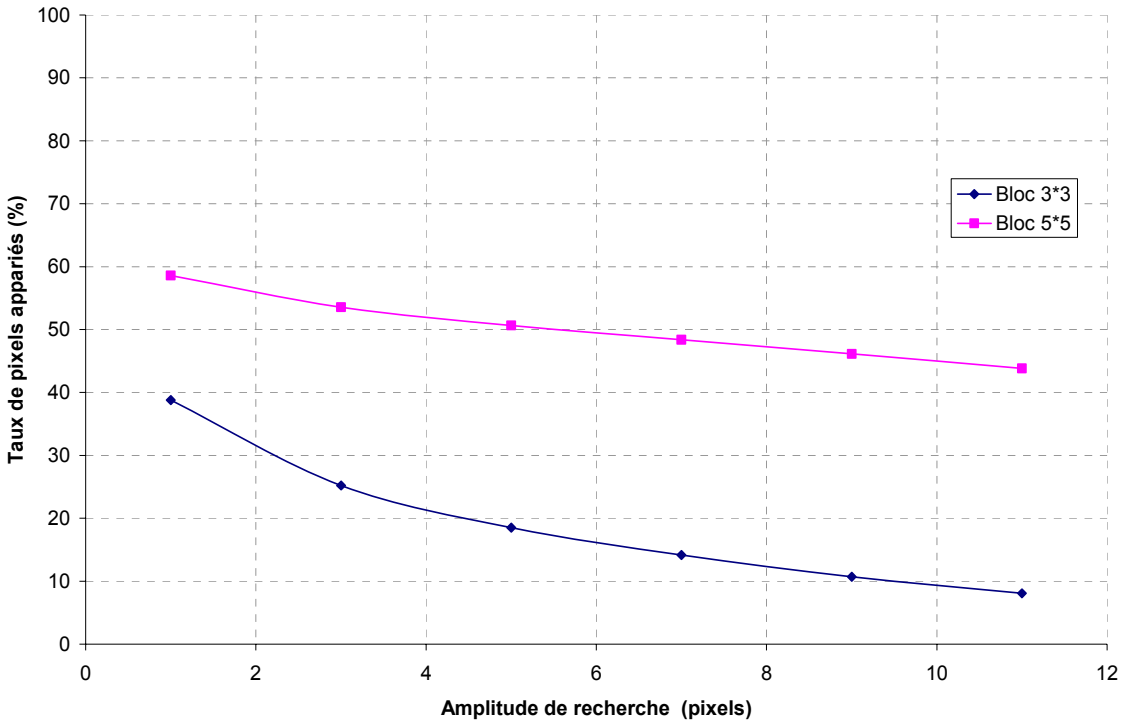


Figure III.18. Taux de pixels appariements en fonction de la longueur de la zone de recherche pour un codage du recensement sur un voisinage 3×3 et 5×5.

Le taux d'appariements est relativement faible puisqu'il est inférieur à 60% pour une amplitude de zone de recherche réduite et un voisinage de codage de 5×5 pixels, pour décroître jusqu'à 45% environ. Ces résultats semblent, a priori, insuffisants pour nous permettre une estimation du mouvement global performante. Le taux d'appariements est encore plus faible et inexploitable dans le cas d'un codage 3×3 pixels.

IV.1.d. Bilan des performances

Les taux d'appariements que nous avons obtenus par nos trois techniques d'estimation des mouvements locaux indiquent que la plus robuste est celle de l'appariement de blocs de pixels, avec un taux voisin de 99%. La taille du bloc apparié est apparue peu influente pour cette technique, de même que l'amplitude de recherche des blocs.

Vient ensuite la technique de l'appariement de codes de texture, avec un taux de pixels appariés voisin de 90% et variant peu avec l'amplitude des mouvements recherchés dans le cas d'un codage 5×5. Dans le cas de codes 3×3, le taux est légèrement inférieur à 80% pour une zone de recherche restreinte à quelques pixels et décroît jusqu'à 55% pour une zone de recherche de 11 pixels, soit 4% de la taille image.

Enfin la technique de l'appariement de bloc de pixels avec extraction de contrastes a montré de faibles performances, puisque le taux d'appariement de blocs 5×5 pixels est inférieur à 60% pour une amplitude de zone de recherche faible (<1% de la taille image) et décroît jusqu'à 45% pour une amplitude de 4% de la taille image.

IV.2. Estimation du mouvement global (EMG)

Nous estimons les mouvements locaux périphériques à l'aide des deux techniques sélectionnées précédemment. Pour chacune d'elles, nous caractérisons la pertinence des mouvements globaux estimés : d'une part en reportant sur un graphique, en coordonnées pixels, la position du centre image au cours la séquence vidéo[72], que nous comparons aux positions réelles ; et d'autre part en étudiant les statistiques des erreurs entre les mouvements globaux estimés et réels.

Les séquences vidéo tests sont celles que nous avons décrites dans la section III. Il s'agit donc de vidéos artificielles et paramétrées, ou réelles. Dans le cas réel, nous associons sur le même graphique les résultats de nos estimations avec celles de l'algorithme de Odobez & Bouthemy afin de mieux apprécier les performances.

Concernant l'étape de pondération (étape 4 sur Figure III.8), nous introduisons un seuil afin de détecter les écarts importants sur les différences des déplacements locaux décrits par notre estimation initiale du modèle (sans pondération) et nos mesures locales des mouvements périphériques. Chacune des mesures locales dont la différence dépasse ce seuil se voit attribuer un poids de confiance faible, de valeur égale à 0.2, au lieu de 1 pour les autres mesures. Ce seuil est défini proportionnellement à l'amplitude du mouvement décrite par la première estimation du modèle de mouvement global. Il a pour valeur 1 ici, c'est à dire qu'il faut une mesure de mouvement local supérieure au double du mouvement local décrit par l'estimation initiale du modèle pour que le poids de la mesure soit affecté[73].

IV.2.a. EMG à partir d'appariements de codes de texture

Nous considérons ici les résultats obtenus pour un codage sur des tailles de bloc de 3×3 et 5×5 pixels, sur la séquence artificielle « Nature ». L'amplitude de la zone de recherche est de 7 pixels, soit une amplitude maximum de mouvements locaux périphériques de près de 3% de la taille image, en accord avec les spécifications que nous avons établies au paragraphe I.1.

[72] *Nous obtenons ces positions en ajoutant les mouvements globaux successivement estimés au cours de la séquence.*

[73] *Ici nous employons une pondération binaire afin de minimiser les ressources de calcul nécessaires, mais les performances de l'algorithme peuvent être améliorées en adoptant une pondération continue.*

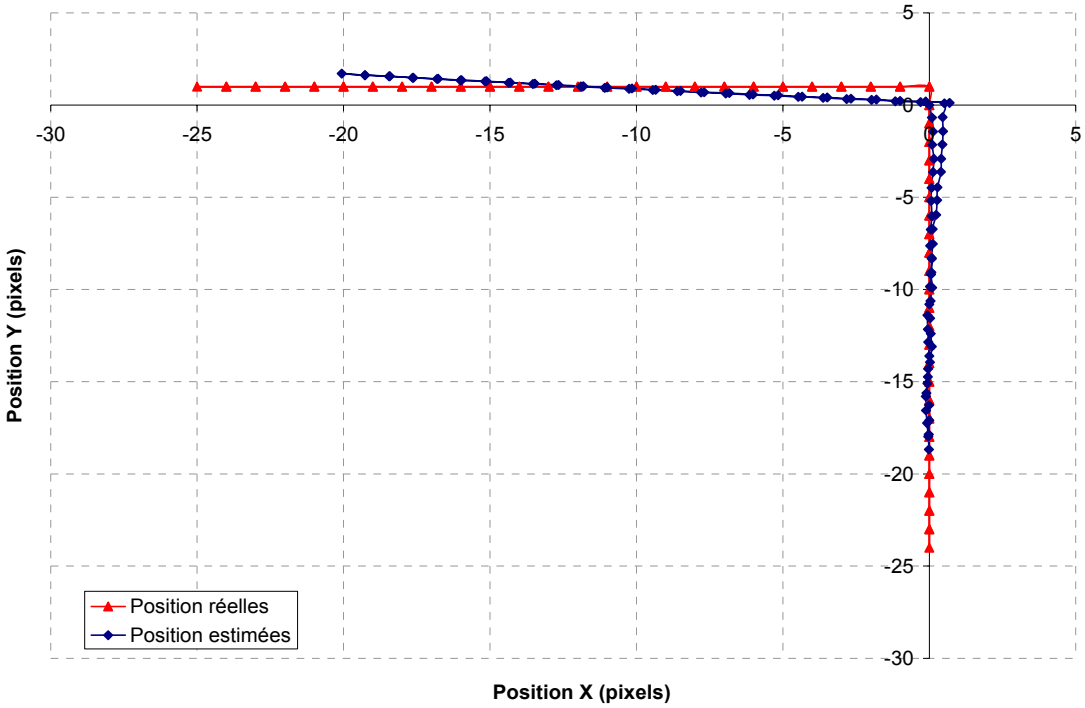


Figure III.19. Dans le plan focal et en coordonnées pixels : positions réelles et estimées (par appariements de codes de texture 3×3) du centre image au cours de la séquence vidéo artificielle et paramétrée « Nature ».

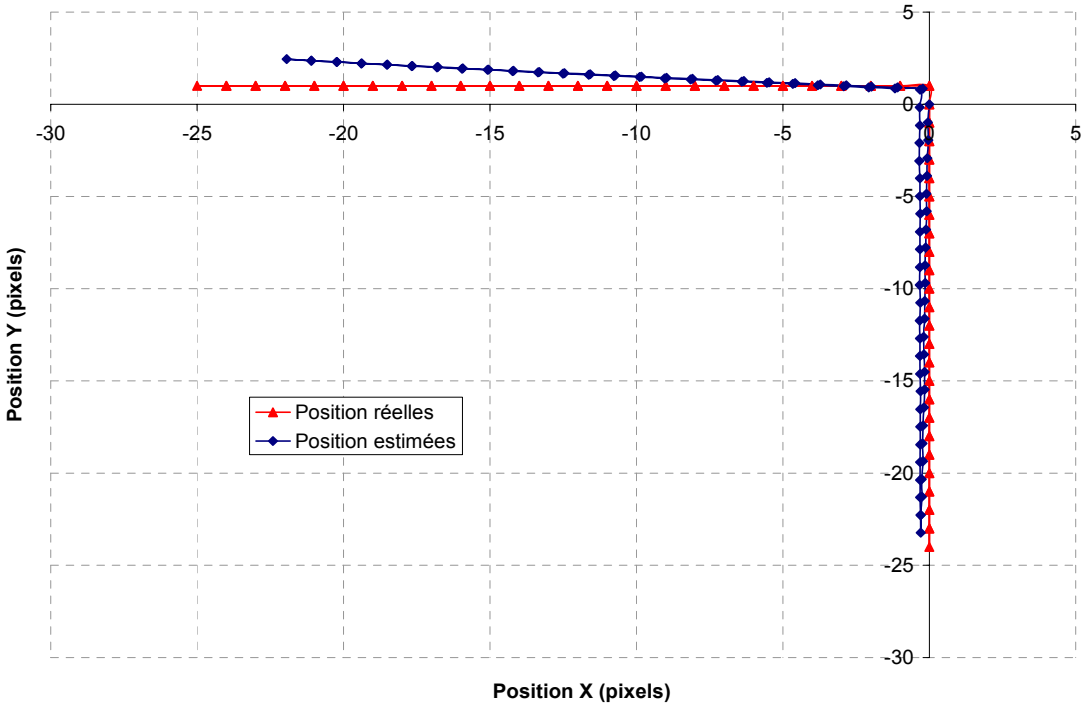


Figure III.20. Dans le plan focal et en coordonnées pixels : positions réelles et estimées (par appariements de codes de texture 5×5) du centre image au cours de la séquence vidéo artificielle et paramétrée « Nature ».

Nous constatons l'apport du codage 5×5, mais il n'est pas suffisant ici pour justifier l'augmentation de complexité qu'il engendre. Les statistiques associées à ces estimations sont reportées dans le tableau ci-dessous :

$\lvert MG_{réel} - MG_{estimé} \rvert$	*Bloc 3×3*		*Bloc 5×5*	
	Tx	Ty	Tx	Ty
Moyenne (%)	10.1	14.2	7.3	5
Ecart type (%)	7.1	9.6	6.2	1.9
Dynamique (%)	20	30	16	7

Tableau III-2. Statistiques sur les différences absolues d'estimation des 4 paramètres du modèle : moyenne, écart type, et dynamique[74].

L'inconvénient des séquences artificielles de caractérisation est que les conditions lumineuses d'acquisition, ainsi que le contenu de la scène sont trop parfaits. En effet, les hypothèses effectuées lors de la conception de notre technique concernant le type de scène et les variations de luminosités sont parfaitement remplies. Ces hypothèses, classiquement employées par la communauté de traitement d'images, sont notamment : une scène plane, avec peu ou pas d'objets en mouvement singuliers, avec de faibles variations locales d'éclairement.

Dans le cas réel maintenant, sur la séquence « Nature » acquise à l'aide de la caméra Canon, nous appliquons l'algorithme avec les mêmes paramètres : taille des blocs de codage de 3×3 et 5×5, avec une amplitude de recherche des codes de 7 pixels (3% de la taille image).

Nous obtenons de piètres résultats ici, qui sont liés à la compression importante des données. En effet, les codes ne sont alors plus appariés correctement à cause du filtrage passe-bas résultant du format de compression ainsi que des discontinuités entre blocs de pixels voisins. On note cependant une amélioration des résultats dans le cas d'un codage 5×5, par rapport au 3×3, mais les résultats ne sont tout de même pas convaincants.

[74] Les paramètres theta et alpha sont vides car nous n'avons pas pu les déduire des paramètres restitués par l'algorithme de Odobez et Bouthemy.

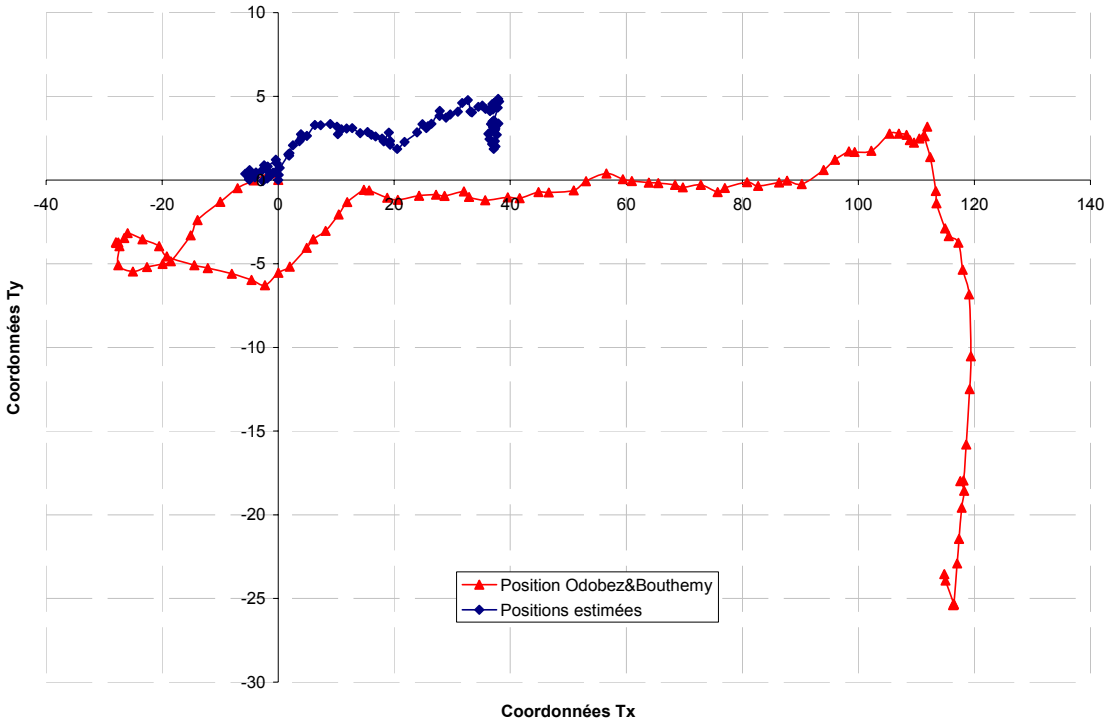


Figure III.21. Dans le plan focal et en coordonnées pixels : positions réelles et estimées (par appariements de codes de texture 3×3) du centre image au cours de la séquence vidéo réelle « Nature » compressée à 320 kbps (en 320×240).

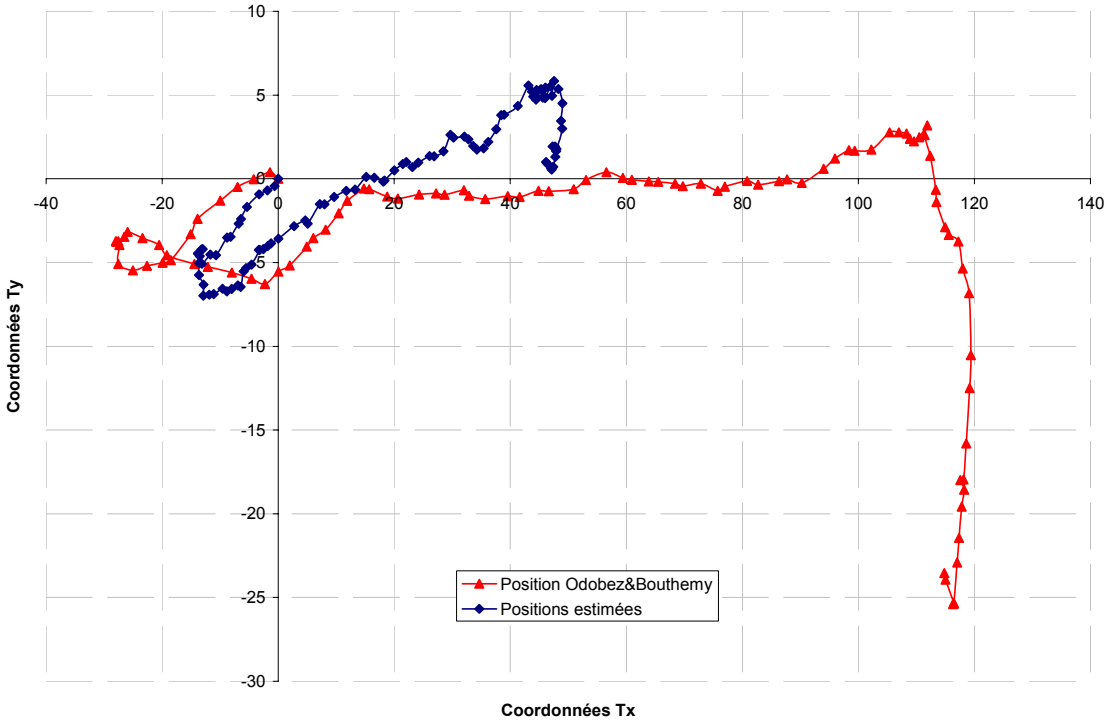


Figure III.22. Dans le plan focal et en coordonnées pixels : positions réelles et estimées (par appariements de codes de texture 5×5) du centre image au cours de la séquence vidéo réelle « Nature » compressée à 320 kbps (en 320×240).

IV.2.b. EMG à partir d'appariements de blocs de pixels

Nous estimons ici les mouvements locaux périphériques par la technique de l'appariement de bloc de pixels, ou « block-matching ». Comme dans le cas précédent, l'amplitude de la zone de recherche est de 7 pixels, soit une amplitude maximum de mouvements locaux périphériques de 3% de la taille des images.

Nous reportons ci-après les résultats obtenus sur la séquence artificielle « Nature », pour des tailles de blocs de 3×3 et 5×5 pixels.

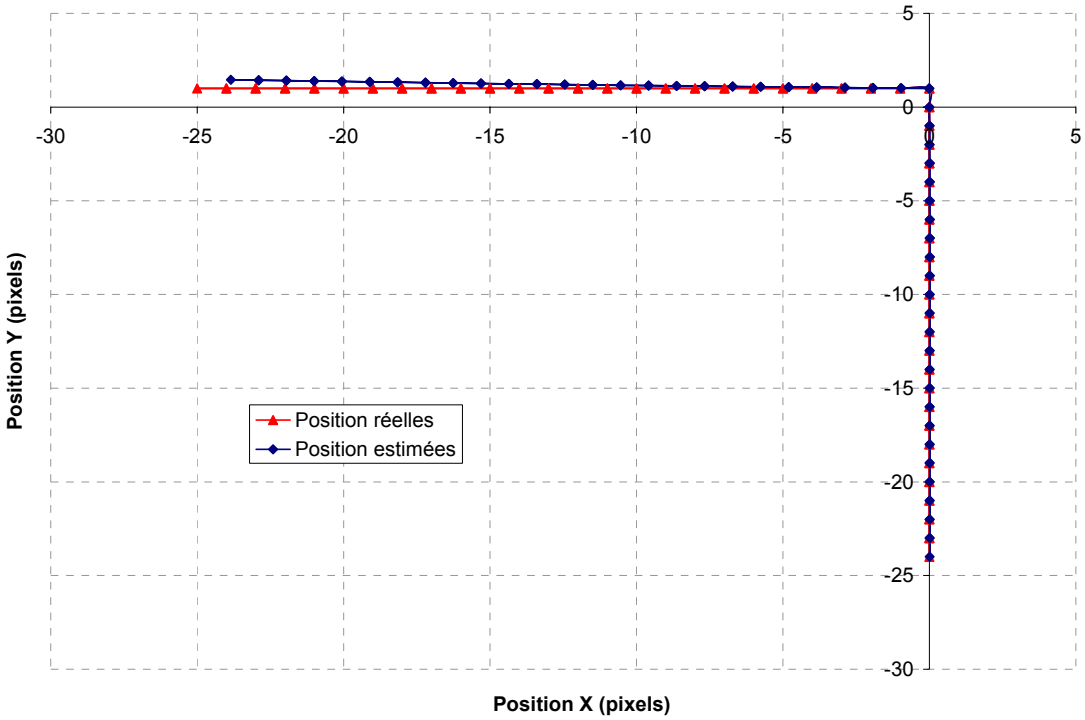


Figure III.23. Dans le plan focal et en coordonnées pixels : positions réelles et estimées (par appariements de blocs de 3×3 pixels) du centre image au cours de la séquence vidéo artificielle et paramétrée « Nature ».

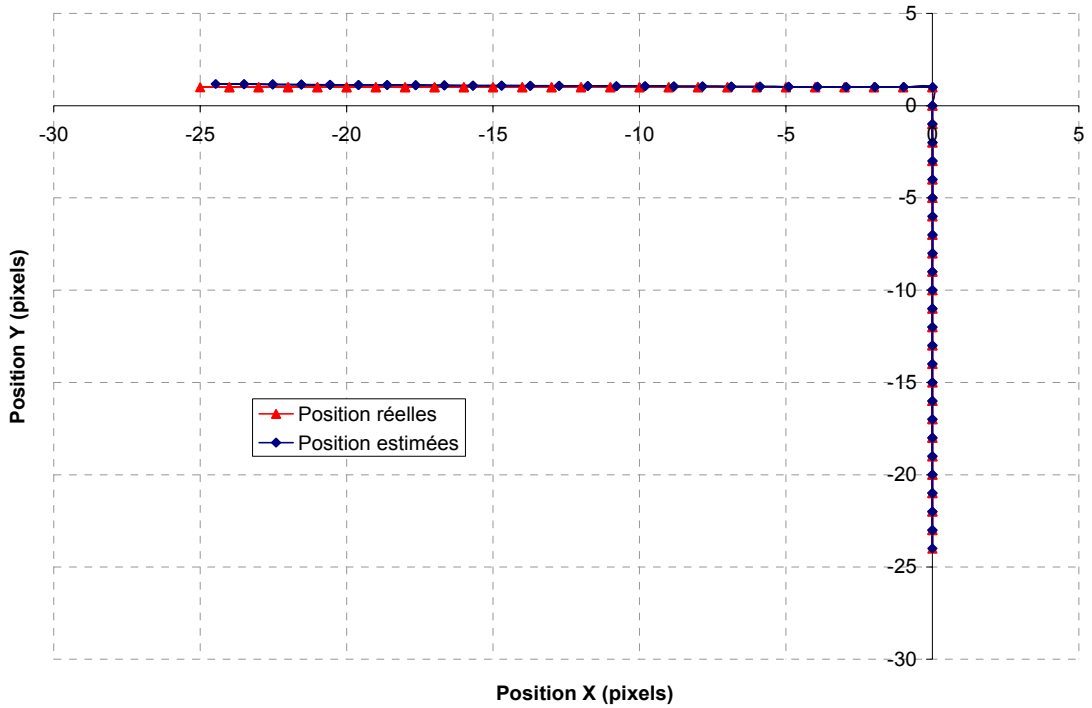


Figure III.24. Dans le plan focal et en coordonnées pixels : positions réelles et estimées (par appariements de blocs de 5×5 pixels) du centre image au cours de la séquence vidéo artificielle et paramétrée « Nature ».

Nous constatons que l'estimation du mouvement global est plus fidèle dans le cas de blocs de taille 5×5, ce qui se confirme lorsque nous étudions les erreurs d'estimations des paramètres du modèle. Nous les reportons ci-dessous, dans le cas de séquences synthétiques :

$\lvert MG_{réel} - MG_{estimé} \rvert$	*Bloc 3×3*		*Bloc 5×5*	
	Tx	Ty	Tx	Ty
Moyenne (%)	3	1	1.9	0.48
Ecart type (%)	0.8	0.6	0.5	0.25
Dynamique (%)	3	1.9	2.1	1.1

Tableau III-3. Statistiques sur les valeurs absolues des différences d'estimation des 4 paramètres du modèle : moyenne, écart type, et dynamique.

Les précisions que nous obtenons sur ces séquences artificielles sont encourageantes puisque les paramètres du modèle sont estimés avec une erreur moyenne en valeur absolue de 1% environ et un écart type voisin de 0.5%. Nous caractérisons maintenant l'algorithme sur séquences réelles.

Le constat que nous avons fait qui consiste à remarquer que la précision d'estimation s'améliore avec des tailles de blocs plus grandes se vérifie aussi dans le cas réel. Nous reportons les estimations obtenues dans le cas de mouvements locaux périphériques extraits sur la séquence « Nature » compressée à 320 kbps (en 320×240), avec des tailles de blocs de 3×3 et 5×5 pixels, et une distance de recherche de 7 pixels (~3% de la taille image). Nous reportons également le cas d'une taille de blocs de 7×7 pixels.

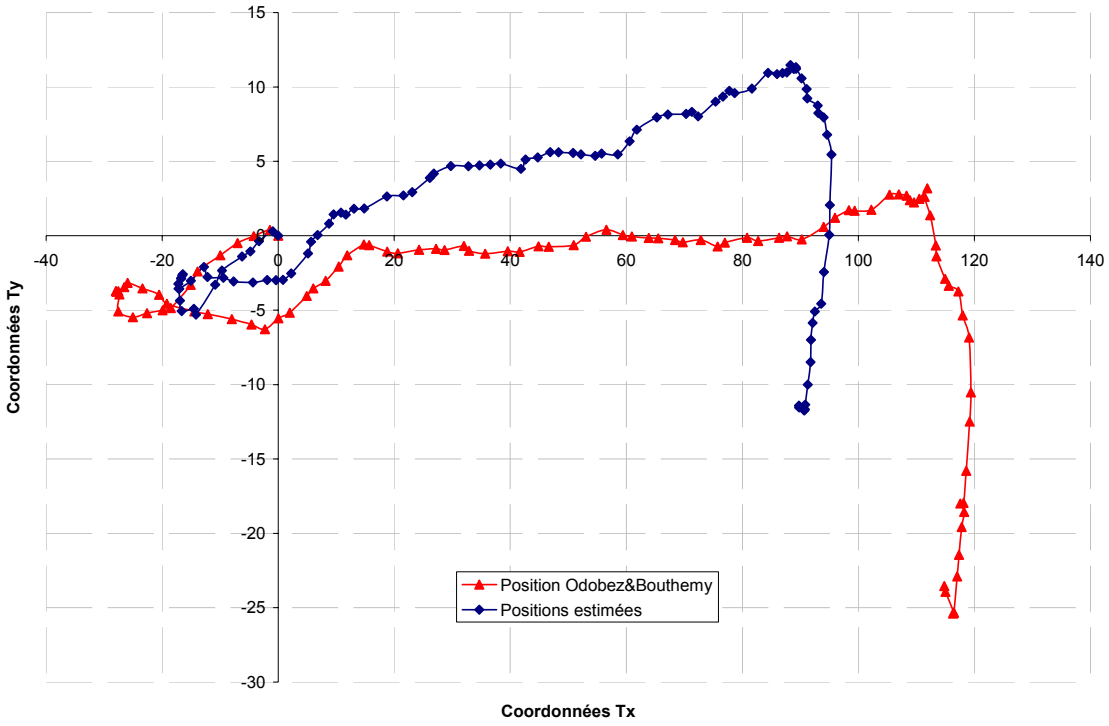


Figure III.25. Dans le plan focal et en coordonnées pixels : positions réelles et estimées (par appariements de blocs de 3×3 pixels) du centre image au cours de la séquence vidéo réelle « Nature » compressée à 320 kbps (en 320×240).

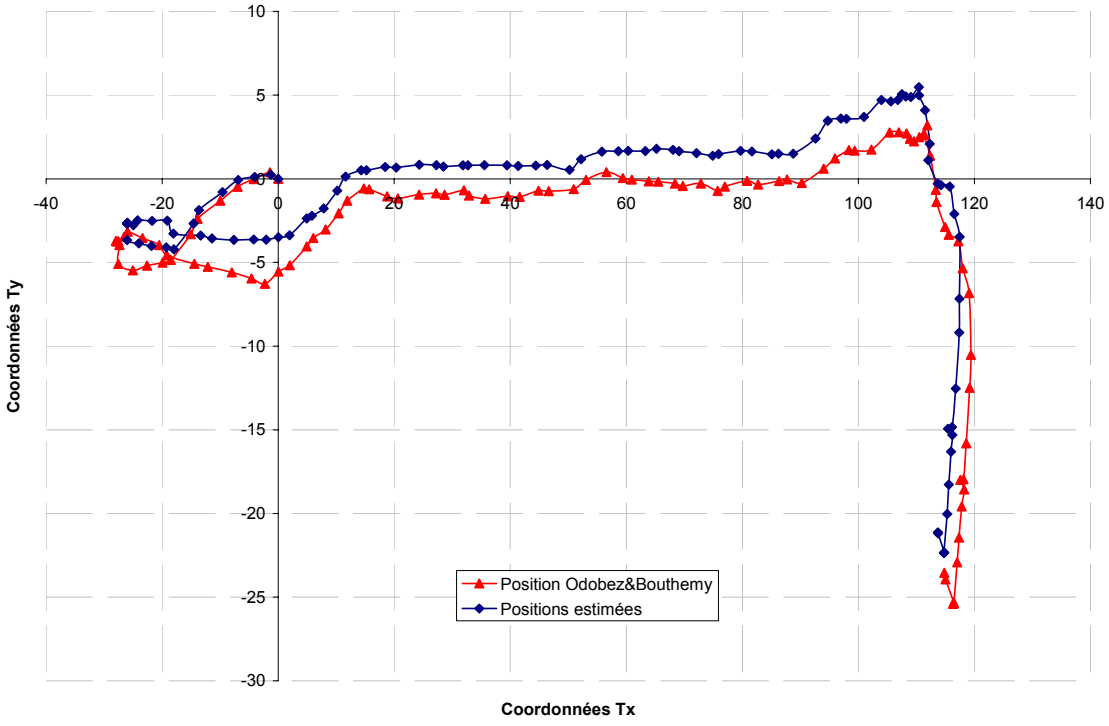


Figure III.26. Dans le plan focal et en coordonnées pixels : positions réelles et estimées (par appariements de blocs de 5×5 pixels) du centre image au cours de la séquence vidéo réelle « Nature » compressée à 320 kbps (en 320×240).

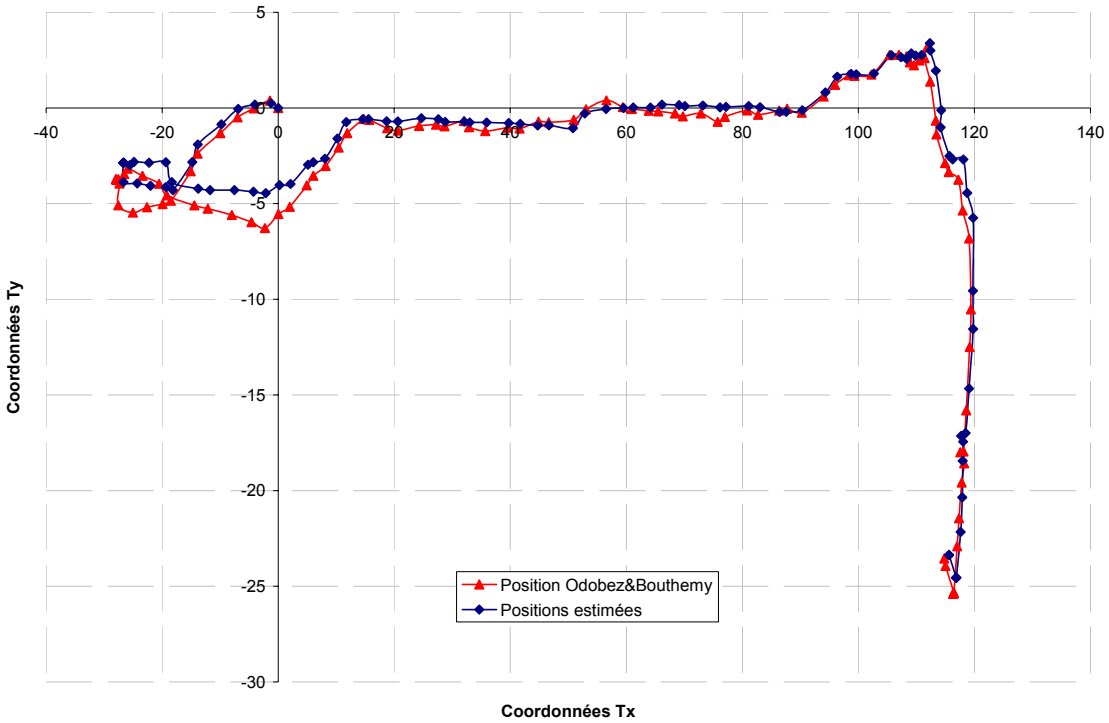


Figure III.27. Dans le plan focal et en coordonnées pixels : positions réelles et estimées (par appariements de blocs de 7×7 pixels) du centre image au cours de la séquence vidéo réelle « Nature » compressée à 320 kbps (en 320×240).

En prenant pour référence les paramètres estimés par la technique de Odobez & Bouthemy, les précisions que nous atteignons sur ces séquences vidéo possèdent les statistiques suivantes :

$\lvert MG_{réel} - MG_{estimé} \rvert$	Bloc 3×3		Bloc 5×5		Bloc 7×7	
	Tx	Ty	Tx	Ty	Tx	Ty
Moyenne (%)	25	12	1.1	2.4	0.8	0.2
Ecart type (%)	71	46	18	23	18	23
Dynamique (%)	435	260	90	124	93	118

Tableau III-4. Statistiques sur les valeurs absolues des différences d'estimation des paramètres du modèle sur la séquence « Nature », compressée à 320 kbps (en 320×240) : moyenne, écart type, et dynamique.

Les meilleures précisions sont celles des plus grandes tailles de blocs de pixels, en étant inférieures à 1% pour un écart type de 18% dans le cas 7×7 pixels. Pour des blocs de 5×5 pixels, la précision se situe autour de quelques pour cent avec un écart type de près de 20%.

Nous pouvons remarquer aussi sur l'ensemble des résultats une sensibilité de l'erreur à la direction du mouvement qui a pour effet de biaiser le mouvement estimé. Ceci est dû à l'ordre d'exploration des blocs de pixels et à un critère de comparaison qui n'est pas suffisamment discriminant dans la procédure d'appariement de blocs de pixels.

IV.2.c. EMG à partir d'appariements d'images contrastées

La troisième technique d'estimation des mouvements locaux périphériques consiste à appliquer la mise en correspondance de blocs de pixels, comme précédemment, sur nos séquences d'images tests auxquelles nous avons appliqué le filtre de Canny pour en extraire les contrastes. Nous avons effectué les caractérisations pour des tailles de blocs de 3×3, 5×5, et 7×7 pixels, avec une amplitude de la zone de recherche de 7 pixels, soit une amplitude maximum de mouvements locaux périphériques de 3% de la taille des images.

Les résultats que nous avons obtenus ne sont pas satisfaisants. En effet, dans le cas des séquences synthétiques, qui est pourtant favorable aux algorithmes d'estimation du mouvement, les positions du centre image au cours de la séquence divergent très rapidement. Nous avons réalisé ces estimations du mouvement global avec plusieurs seuils d'extraction des contrastes, mais les résultats étaient là aussi peu concluants.

Pourtant, si nous considérons une partie de la séquence vidéo « Pano_ext » (images 100 à 199 par exemple), à laquelle nous détectons les contrastes de façon unidimensionnelle avec un filtre de Sobel horizontal[75], et que nous extrayons et juxtaposons les lignes 5 et 235 de la séquence d'images (cf. Figure III.28), nous visualisons bien le mouvement de translation diagonal de droite à gauche et de bas vers haut de la caméra. Cela nous prouve donc deux choses : d'une part qu'il est bien possible de percevoir le mouvement global inter images à partir d'une détection de contraste, et d'autre part que nous devons repenser notre algorithme d'estimation des mouvements locaux périphériques.

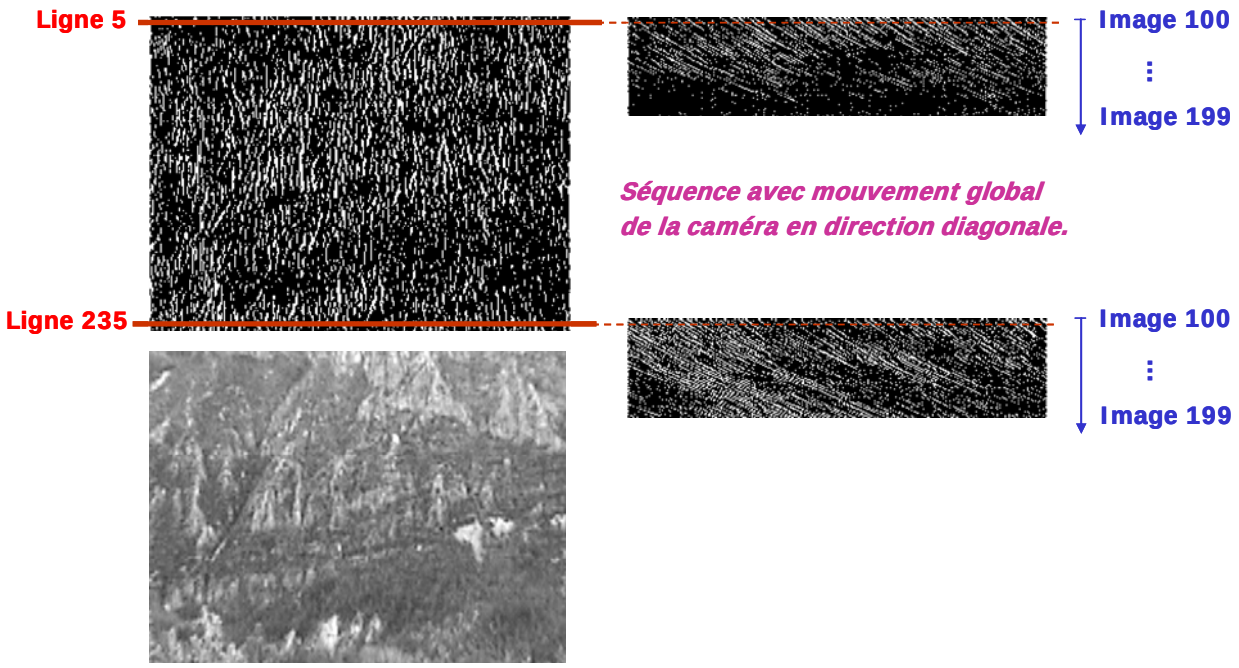


Figure III.28. Extraction des lignes 5 et 235 d'une partie de la séquence vidéo « Pano_ext » (images 100 à 199) avec détection des contrastes verticaux.

[75] *C'est-à-dire une détection des contrastes verticaux.*

IV.3. Précision de l'E.M.G. pour la stabilisation vidéo

Nous venons d'étudier la précision et le comportement de nos algorithmes d'estimations du mouvement global au cours de séquences vidéo. Cependant, comme nous l'avons vu au paragraphe § I.2. , la stabilisation vidéo implique de mettre en œuvre un filtrage afin de discerner le mouvement intentionnel du mouvement parasite, à compenser. Ainsi, nous représentons sur le graphique ci-dessous l'évolution du mouvement à compenser CX au cours de la séquence vidéo « Nature ».

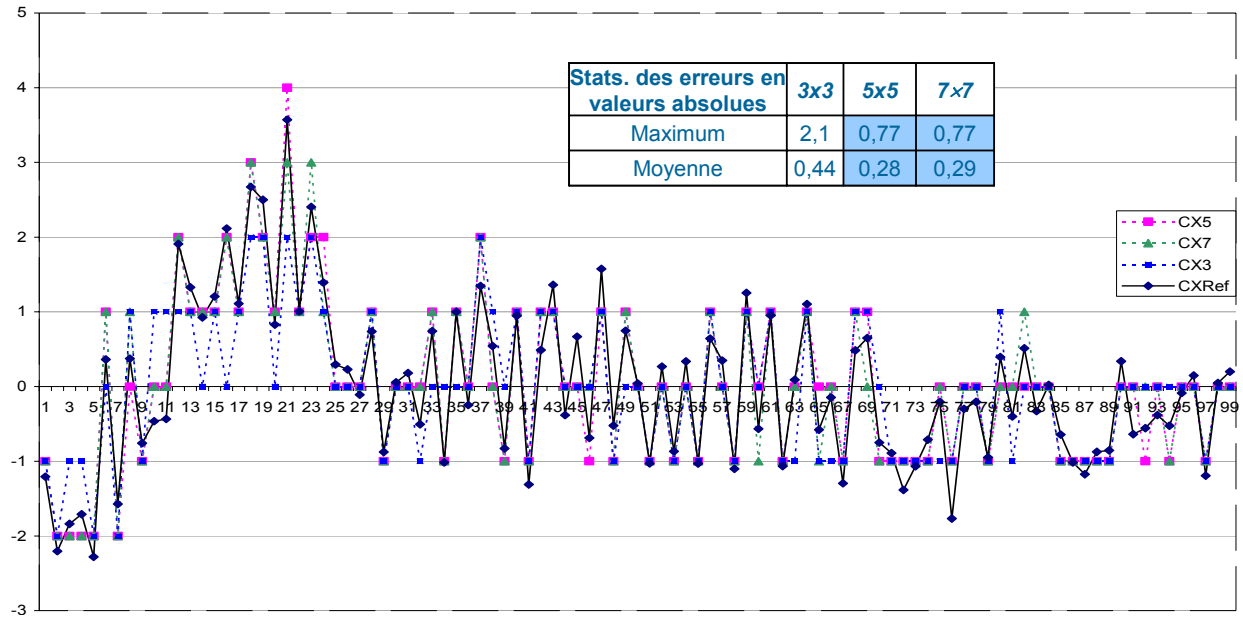


Figure III.29. Evolution de l'amplitude du mouvement à compenser CX au cours de la séquence "Nature". Les mouvements locaux périphériques sont estimés ici par appariement de blocs de pixels (de tailles 3x3, 5x5, ou 7x7), et le poids du filtre α est de 0.15 .

Nous remarquons que les erreurs obtenues par différence entre le mouvement à compenser théorique CXRef (basé sur l'EMG de Odobez-Bouthemy) et celui résultant de notre technique CX sont toujours inférieures au pixel dès lors que l'on considère des blocs de pixels 5x5 ou 7x7 (maximum de valeur 0.77, cf. Figure III.29).

Etant donné que la stabilisation vidéo que nous envisageons est basée sur un recadrage des images au pixel près, nous en concluons que notre technique d'EMG est satisfaisante pour cette application.

CONCLUSION

Dans ce chapitre, nous avons tout d'abord établi les spécifications de notre système d'estimation du mouvement global en vue de la stabilisation vidéo électronique. La compensation du mouvement indésiré s'effectue par recadrage, ainsi ce sont les mouvements globaux de translation que nous corrigeons en priorité. A la cadence de 25 images par secondes, les amplitudes maximales de mouvements que nous avons décidé de considérer sont inférieures à 3% de la taille image.

Nous avons ensuite mis en œuvre une technique d'estimation du mouvement global basée sur la mesure de mouvements locaux en bord d'image. Nous avons identifié trois méthodes pour déterminer ces mouvements locaux périphériques, qui appartiennent toutes à la classe des techniques de mise en correspondance. Nous estimons les paramètres du modèle de mouvement global par un processus d'optimisation au sens des moindres carrés pondérés, nous permettant d'obtenir les quatre paramètres du modèle de mouvement que nous avons choisi. Ce modèle est à quatre paramètres afin de permettre d'éventuelles améliorations de la stabilisation, notamment en rotation et en zoom.

En appliquant l'algorithme ainsi développé à des séquences vidéo artificielles et réelles, nous obtenons dans le pire cas (séquences très compressées) une précision sur le mouvement global estimé voisine de 1% en moyenne, avec un écart type proche de 20%. Ceci en estimant les mouvements locaux à l'aide de la mise en correspondance de blocs de pixels de taille 5×5.

Nous avons également étudié l'évolution de l'amplitude du mouvement à compenser au cours de plusieurs séquences vidéo. Ceci pour montrer que l'erreur que nous commettons sur ce mouvement (par rapport à une compensation mettant en œuvre une EMG Odobez-Bouhemy) reste toujours inférieure au pixel. Le recadrage des images suivant ce mouvement à compenser s'effectuant au pixel près, cela atteste de la faisabilité et la pertinence de notre approche d'estimation du mouvement global, ce qui nous conduit à évoquer son intégration sur silicium.

Chapitre IV.

INTEGRATION A UN IMAGEUR CMOS

INTRODUCTION

Ayant validé de manière logicielle notre technique d'estimation du mouvement global par mesures périphériques. Nous nous intéressons dans ce chapitre à l'intégration de cette technique à un imageur CMOS actuellement fabriqué par la société STMicroelectronics, en privilégiant un traitement du signal par approche « rétine », c'est-à-dire au plus près de la phototransduction, au niveau pixel.

La première partie est dédiée à une étude de la charge de calcul en vue d'une intégration sur architecture numérique « classique » de l'algorithme développé. Nous différencions pour cela chaque technique de détermination des mouvements locaux périphériques.

Ensuite, dans un esprit d'Adéquation-Algorithme-Architecture, nous discutons des différentes architectures systèmes envisageables pour intégrer cette estimation du mouvement global à un imageur. Nous considérons alors le partitionnement du traitement : dans le plan focal et en post-traitement, et nous dégageons l'architecture système retenue. Celle-ci associe ces deux types de traitement afin d'accomplir les deux tâches de restitution de la vidéo et d'estimation du mouvement global de manière optimisée.

Cette architecture met en œuvre un traitement au niveau pixel pour faciliter les mesures de mouvements périphériques. Le bruit présent à ce niveau de la chaîne de traitement du signal doit alors être pris en compte car, par souci de minimisation de la surface pixel, aucune réduction du bruit ne peut être intégrée à ce niveau. Nous caractérisons donc son influence sur les traitements d'images envisagés lors du chapitre précédent.

Puis nous décrivons ensuite l'intégration en technologie CMOS de l'ensemble de la chaîne de traitement du signal électrique que nous avons établie. C'est-à-dire d'une part les traitements au niveau pixel, essentiellement destinés aux mesures locales périphériques du mouvement. Mais aussi les architectures numériques spécifiques que nous avons établies pour effectuer certaines tâches de plus haut niveau, en post-traitement.

L'ajout de la fonctionnalité « estimation du mouvement » en temps réel vidéo (inter trame de 33ms) ne doit pas se faire au détriment de la qualité vidéo et doit considérer la contrainte de coût imposée par le marché actuel des imageurs CMOS mégapixel[76] pour la téléphonie mobile. Nous évoquons les aspects « d'Adéquation Algorithme Architecture Application » dans notre dernière partie, où nous évaluons les techniques d'estimation du mouvement global que nous avons proposées en considérant certaines mesures de coût telles que la surface silicium et les ressources matérielles requises, ainsi que la performance comme la précision du mouvement.

[76] *En grand volume, le prix d'un imageur megapixels se situe autour de 10$, et celui d'un capteur « souris optique » est de 1$ environ.*

I. IMAGEUR CMOS ET ESTIMATION DU MOUVEMENT GLOBAL EMBARQUEE

L'objectif de cette première partie est de définir l'architecture système optimale qui va intégrer la fonction de stabilisation vidéo dans un imageur, et plus particulièrement l'estimation du mouvement global, qui requiert généralement plus de la moitié des ressources de calcul totales.

I.1. Evaluation des ressources requises pour l'EMG

Nous évaluons dans cette section la charge de calcul que requiert notre technique d'estimation du mouvement global. Nous nous plaçons pour cela dans le cadre d'une implémentation de l'algorithme sur un processeur. Nous considérons alors qu'une multiplication (notée « mult »), une addition (« add »), une différence (« diff »), ou une division (« div ») sont équivalentes à une opération élémentaire (notée « op »). De même, nous définissons le calcul de la distance de Hamming entre deux nombres, équivalente à un « ouexclusif » et un comptage des bits à « 1 », comme étant, lui aussi, équivalent à une opération.

De plus, nous appelons « S » l'amplitude de recherche lors de nos estimations des mouvements locaux par mise en correspondance. La zone de recherche est alors un carré de coté « 2.S+1 » pixels. Nous notons « N » le nombre total de mesures de mouvements locaux périphériques. « N » est tel que : « N = 2.Nx + 2.Ny », où « Nx » est le nombre de mesures sur chacun des bords haut et bas, et « Ny » le nombre de mesures sur chacun des bords droit et gauche.

I.1.a. Spécifications

Comme nous l'avons décrit en détails lors de la partie II.3. du chapitre précédent, notre démarche pour estimer les 4 paramètres du modèle du mouvement inter trames est la suivante :

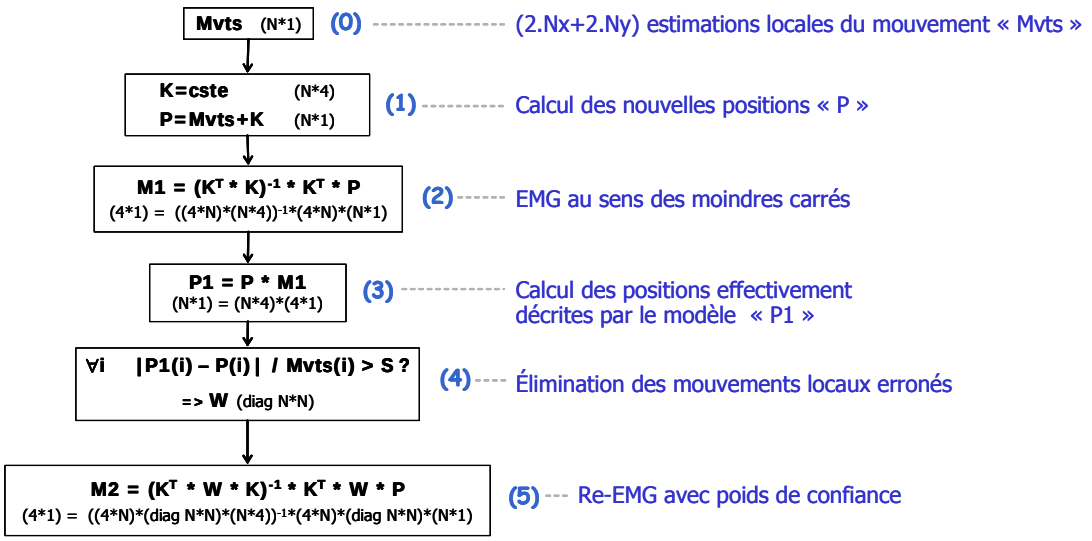


Figure IV.1. Procédure d'estimation du mouvement global inter trames.

Les données d'entrées sont soit les images acquises, dans le cas d'un traitement « classique », soit les vecteurs mouvements locaux périphériques (Mvts), si le capteur est muni de senseurs de perception des mouvements locaux. Les positions initiales de ces vecteurs mouvements, que l'on appelle les points d'intérêts périphériques, sont quant à elles connues et communes pour les deux approches.

La cadence de traitement à respecter pour accomplir la stabilisation vidéo est le temps réel vidéo, c'est-à-dire un temps inter trames vidéo, donc 33 ms au maximum.

On s'aperçoit d'après la Figure IV.1 que les opérations mises en jeu dans ce traitement sont des sommes, des produits, des transpositions, et des inversions de matrices. Nous analysons maintenant en détail le coût de calcul de ces opérations.

I.1.b. Détermination de la charge de calcul

Pour évaluer cette charge de calcul totale, nous parcourons pas à pas l'algorithme illustré Figure IV.1. Nous déterminons à chaque étape le nombre d'opérations élémentaires à réaliser.

Le terme « opération élémentaire » que nous employons ici désigne une addition, une différence, une multiplication, une division, ou une comparaison.

- Etape 0 : mouvements locaux périphériques « Mvts ».

Dans le cas où ces mouvements locaux ne sont pas restitués par le capteur[77], l'étape 0 consiste à les estimer. Pour cela, nous considérons deux techniques que nous avons présentées dans le chapitre précédent, toutes deux basées sur la mise en correspondance de pixels entre deux images. Pour la première nous mettons en œuvre la transformée du recensement, et pour la seconde il s'agit de considérer l'appariement de blocs de pixels.

→ Mise en correspondance des transformées du recensement.

Rappelons que le principe est ici de créer, pour chaque pixel de l'image, un code issu de la concaténation des résultats binaires des comparaisons de sa luminance avec celle de ses voisins (cf. Figure IV.2). Cette transformée est effectuée une seule fois pour chaque pixel (à la différence des techniques de « block matching »).

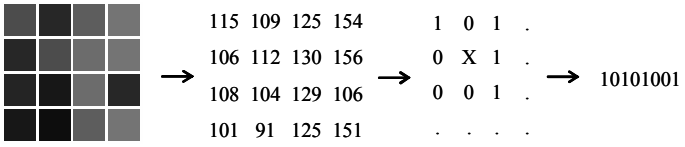


Figure IV.2. Codage du recensement pour un voisinage 3×3.

[77] *C'est-à-dire qu'ils ne sont pas mesurés au niveau pixel, à la différence de ce que nous proposons dans la partie III. de ce chapitre.*

Ainsi, en considérant un codage sur un voisinage « M×M », un code est obtenu à l'aide de « M²-1 » tests d'entiers. Soit « q1 » le nombre d'opérations nécessaires pour effectuer la transformée du recensement sur des voisinages de « M×M » dans une bande de largeur 2S+1 pixels à la périphérie de l'image, « q1 » vaut, pour une image de taille Ny × Nx :

$q1 = [2 \times Ny.(2.S+1) + 2 . (Nx-4.S)\times(2.S+1)] . (M^2-1)$ tests,
soit encore : $q1 \sim 2.(2.S+1) . (Ny+Nx) . (M^2-1)$ op.

Le Tableau IV.1 ci-dessous fournit quelques exemples du nombre d'opérations mises en jeu pour réaliser la transformée du recensement sur la périphérie d'une image 800×600, pour un voisinage de codage de « M×M » pixels et une amplitude de déplacements recherchés de « S » pixels :

M×M ↓ | S →	10	15	20	25	30
3×3	470 kop.	694 kop.	918 kop.	1.14 Mop.	1.3 Mop.
5×5	1.4 Mop.	2.1 Mop.	2.8 Mop.	3.4 Mop.	4.1 Mop.
8×8	3.7 Mop.	5.5 Mop.	7.2 Mop.	9 Mop.	10.8 Mop.

Tableau IV.1. Charge de calcul associée à la transformée du recensement sur la périphérie d'une image 800×600.

Un pixel est « reconnu » d'une image à une autre si son code de transformée est identique. Ainsi, nous recherchons la distance de Hamming minimum entre le code initial dans l'image 1 et les codes des pixels voisins dans l'image 2, ceci sur une zone carrée de côté « 2.S » pixels. Le nombre « q2 » d'opérations mises en jeu vaut alors :

$q2 = (1\ DH) . (2.S+1)^2 \sim (2.S+1)^2$ op.

Le nombre total d'opérations « Q1 » à réaliser pour réaliser « N » estimations locales de mouvements, en considérant un voisinage de codage de taille « M×M » pixels et sur une zone de recherche de « S » pixels, est égal à la somme des charges de calcul, soit :

Eq. IV-1. $\mathbf{Q1 =} q1 + N.q2 = \mathbf{2.(2.S+1) . (Ny+Nx) . (M^2-1) + N . (2.S+1)^2}$ op.

Le tableau ci-dessous résume les charges de calcul requises en fonction du voisinage de codage et de l'amplitude de recherche des codes, pour 280 mouvements locaux périphériques :

M×M ↓ | S →	10	15	20	25	30
3×3	593 kop.	963 kop.	1.3 Mop.	1.87 Mop.	2.4 Mop.
5×5	1.5 Mop.	2.35 Mop.	3.2 Mop.	4.1 Mop.	5.1 Mop.
8×8	3.8 Mop.	5.7 Mop.	7.7 Mop.	9.7 Mop.	11.8 Mop.

Tableau IV.2. Charge de calcul associée à l'estimation des mouvements locaux périphériques par recensement sur une image 800×600, en considérant 280 mouvements locaux (chaque 10 pixels).

➙ *Full Search Block Matching (FSBM).*

Pour la technique du FSBM, nous considérons la somme des différences absolues (SAD) comme mesure de corrélation, des blocs de taille « M×M » pixels, et une amplitude de recherche de « S » pixels[78]. Chaque mouvement local est alors obtenu en déterminant le minimum de la mesure de dissemblance entre un bloc d'une image et un autre bloc de l'image suivante. Cela met en jeu le nombre « q3 » d'opérations élémentaires valant :

$$q3 = [M.M \text{ diff.} + (M.M-1) \text{ add.} + 1 \text{ test}] \times (2.S+1)^2 \sim 8.M^2.S^2 \text{ op.}$$

Finalement, le nombre « Q2 » total d'opérations pour obtenir « N » mouvements locaux est alors :

Eq. IV-2. **Q2 =** q3 . N = **$8.M^2.S^2.N$ op.**

En estimant un mouvement local chaque dix pixels d'un imageur comprenant 800×600 pixels, soit 280 mouvements locaux, nous obtenons les charges de calcul suivantes :

M×M ↓ S →	10	15	20	25	30
3×3	2 Mop.	4.5 Mop.	8 Mop.	12.6 Mop.	18.1 Mop.
5×5	5.6 Mop.	12.6 Mop.	22.4 Mop.	35 Mop.	50.4 Mop.
8×8	14.2 Mop.	32.2 Mop.	57.1 Mop.	89 Mop.	128 Mop.

Tableau IV.3. Charge de calcul associée à l'estimation des mouvements locaux périphériques par FSBM sur une image 800×600, en considérant 280 mouvements locaux (chaque 10 pixels de l'image).

La charge de calcul requise, dans le cas qui nous concerne d'une amplitude de recherche de 20 pixels[79] et pour estimer les 280 mouvements locaux par la technique de corrélation des codes d'une transformée du recensement sur un voisinage 5×5 est environ **7 fois plus faible** que par corrélation des blocs de pixels par technique du FSBM.

- Etape 1 : calcul des nouvelles positions « P ».

A partir des « N » coordonnées d'origine des vecteurs mouvements locaux, et des vecteurs mouvements précédemment estimés « Mvt », on calcule les nouvelles coordonnées des positions « P » par sommation :

$$\forall i \in [1;N],\ P(i) = K(i,1) + Mvt(i).$$

Le nombre d'opérations correspondantes est : N additions ~ **N op. = Q3**

[78] *C'est-à-dire que nous considérons une amplitude de déplacement du bloc initial d'au plus « S » pixels entre deux images.*

[79] *20 pixels représentent 3% de la taille d'une image 800×600.*

- Etape 2 : Première estimation du modèle de mouvement global « M1 ».

$$\boldsymbol{P} = \begin{pmatrix} X_{1,1}(t+dt) \\ \dots \\ X_{1,Nx}(t+dt) \\ X_{Ny,1}(t+dt) \\ \dots \\ X_{Ny,Nx}(t+dt) \\ Y_{1,1}(t+dt) \\ \dots \\ Y_{Ny,1}(t+dt) \\ Y_{1,Nx}(t+dt) \\ \dots \\ Y_{Ny,Nx}(t+dt) \end{pmatrix} = \boldsymbol{K} \begin{pmatrix} X_{1,1}(t) & Y_{1,1}(t) & 1 & 0 \\ \dots & \dots & \dots & \dots \\ X_{1,Nx}(t) & Y_{1,Nx}(t) & \dots & \dots \\ X_{Ny,1}(t) & Y_{Ny,1}(t) & \dots & \dots \\ \dots & \dots & \dots & \dots \\ X_{Ny,Nx}(t) & Y_{Ny,Nx}(t) & 1 & 0 \\ Y_{1,1}(t) & X_{1,1}(t) & 0 & 1 \\ \dots & \dots & \dots & \dots \\ Y_{Ny,1}(t) & X_{Ny,1}(t) & \dots & \dots \\ Y_{1,Nx}(t) & X_{1,Nx}(t) & \dots & \dots \\ \dots & \dots & \dots & \dots \\ Y_{Ny,Nx}(t) & X_{Ny,Nx}(t) & 0 & 1 \end{pmatrix} \times \boldsymbol{M} \begin{pmatrix} \alpha\cos\theta \\ \alpha\mathrm{Sin}\theta \\ Tx \\ Ty \end{pmatrix}$$

Figure IV.3. Système linéaire d'équations.

Au chapitre précédent, nous avons obtenu le système d'équation surdéterminé ci-dessus à partir des deux équations décrivant la transformation géométrique inter images. La matrice « K » contient les positions initiales des pixels et lie la matrice « P » de leurs positions finales dans l'image suivante au modèle de mouvement global « M ». Cette matrice « K » est constante, le terme « $(K^T \times K)^{-1} \times K^T$ » est alors une matrice 4×N constante qui peut être calculée a priori.

Le calcul à réaliser pour estimer le modèle de mouvement global au sens des moindres carrés est le suivant : « $(K^T \times K)^{-1} \times K^T \times P$ ». Ce produit est un produit de matrices (4×N)×(N×1), la charge de calcul est donc de :

4×(Nmult. + (N-1) add.) ~ **8N-4 op. = Q4**

- Etape 3 : Calcul des positions effectivement décrites par le modèle.

Il s'agit du produit de matrices « P1 = K × M1 », de taille (N×4)×(4×1). Etant donné que la matrice « K » contient au moins un zéro par ligne, le nombre d'opérations est :

(3 mult + 2 add.) * N = 5 N op. = Q5

- Etape 4 : Adéquation du modèle global aux mouvements locaux et création de la matrice W.

Il s'agit de vérifier que le mouvement local soit en accord avec le mouvement global estimé précédemment. C'est-à-dire que nous nous assurons que ce mouvement local ne correspond pas au mouvement parasite d'un objet mobile dans la scène, qui fausserait notre estimation. Pour cela, nous introduisons un seuil que le rapport du mouvement local mesuré par celui décrit par le modèle du mouvement global précédemment estimé ne doit pas dépasser. Si ce seuil est franchi, nous attribuons un poids faible à la mesure locale du mouvement. Le seuil et le poids sont des coefficients réels positifs de valeurs respectives 1 et 0.2. Nous avons défini ces valeurs car elles nous ont permis d'obtenir les meilleures estimations du mouvement global lors de notre série de tests.

Les opérations mises en jeu sont les suivantes :

$$\forall i \in [1;N], \quad Si \ (\ |P1(i)-P0(i)| > |Seuil \times M0(i)|\ ) \ alors \ W(i;i) = poids$$

C'est-à-dire : **N×(1 test. + 1 add. + 1 mult.) ~ 3N op. = Q6**

- Etape 5 : 2e estimation du modèle de mouvement global.

Les calculs à réaliser pour estimer au sens des moindres carrés le modèle de mouvement global sont « $(K^T \times W \times K)^{-1} \times K^T \times W \times P0$ », c'est-à-dire :

1>> « $K^T \times W \times K$ »,

W étant une matrice diagonale, il s'agit d'un produit de matrices (4×N)×diag(N×N)×(N×4). En tenant compte là aussi des particularités de la matrice « K », le nombre d'opérations nécessaires est :

(2×N + 6×N) mult. + (4×(N-1)+10×(N/2-1)) add. = (8×N) mult. + (9×N-14) add. ~ 17N-14 op. = q4

2>> « $(K^T \times W \times K)^{-1}$ »,

Ce calcul est une inversion de matrice (4×4) :

$$\frac{1}{\det(4\times 4)} \times (co\text{-}mat(4\times 4))^T$$

D'où la charge de calcul : 1 div. + 144 mult. + 80 add. ~ 225 op. = q5

3>> « $(K^T \times W \times K)^{-1} \times K^T \times W$ »,

Il s'agit du produit de matrices (4×4)×(4×N)×diag(N×N), avec les simplifications liées à la matrice « K » :

(2×N+3×N) mult. + 2×N add. ~ 7N op. = q6

4>> « $(K^T \times W \times K)^{-1} \times K^T \times W \times P0$ »,

Il ne reste à ce niveau que le produit de matrices (4×N)×(N×1), d'où la charge de calcul suivante :

3×N mult. + (2*(N-1)+2*(N/2-1)) add. = 3×N mult. + (3×N-4) add. ~ 6N-4 op. = q7

Enfin, la charge de calcul totale pour réaliser la deuxième estimation du modèle est la somme des charges de calcul ci'dessus :

Q = q4 + q5 + q6 + q7 = 1 div. + (16×N+144) mult. + (14×N+62) add. ~ 30N+ 207 op.

I.1.c. Charge de calcul totale

Nous synthétisons dans le tableau suivant l'ensemble des charges de calcul nécessaires à l'extraction des 4 paramètres du mouvement global inter trames :

Estim. mvts. locaux (FSBM)	$8 \times M^2 \times N \times S^2$
Estim. mvts. locaux (recens.)	$2 \times (2 \times S+1) \times (Ny+Nx) \times (M^2-1) + N \times (2 \times S+1)^2$
GME	42N + 207

Tableau IV.4. Charge de calcul totale d'estimation du modèle de mouvement global inter trames.

Exemple pratique :

Pour une séquence vidéo SVGA (800×600), avec une amplitude de mouvement de 20 pixels (3% de la taille image), un nombre N = 280 mesures de mouvements locaux périphériques et une taille de blocs de M×M pixels, nous obtenons les répartitions de la charge de calcul ci-dessous :

Estim. Mvts. Périphériques (FSBM, 5×5)	22 400 000 op.	99.95 % de la charge de calcul totale
Estim. Mvts. Périphériques (recens. 5×5)	3 200 000 op.	99.6 % de la charge de calcul totale
Estim. Mvts. Périphériques (recens. 3×3)	1 300 000 op.	99 % de la charge de calcul totale
Est. Mvt. Global	11 967 op.	

Tableau IV.5. Exemple de charge de calcul pour N=208 mouvements locaux périphériques déterminés par corrélation de blocs de pixels et corrélation de transformée du recensement, avec une amplitude de recherche de +/- 16 pixels (3% de la taille image).

Au vu de ces résultats, on remarque que l'essentiel de la charge de calcul est liée à la détermination des mouvements locaux périphériques, s'élevant à plus de 99% de la charge totale. Le processus d'estimation du mouvement global connaissant ces mouvements périphériques ne représente alors qu'une très faible part de la charge de calcul totale (moins de 1%).

I.2. Trois architectures systèmes pour réaliser ce traitement du signal

Nous venons d'étudier, sans aucune considération architecturale, la charge de calcul associée au traitement du signal à mettre en place pour estimer le mouvement global inter trames d'une vidéo fournie par un imageur. Nous nous intéressons dans cette partie à son intégration, d'un point de vue architecture système de traitement du signal.

Dans une architecture d'imageur CMOS actuel, que nous avons présenté au début de ce manuscrit, les pixels ont pour fonction de mesurer la quantité la lumière reçue pendant un intervalle de temps donné. Dans le cas présent, nous nous intéressons à une application spécifique : la perception du mouvement, et nous avons relevé dans notre état de l'art l'existence de pixels dédiés à cette tâche. Ces pixels sont potentiellement intéressants, cependant leurs architectures et leurs fonctionnements sont très différents de celui des pixels « image », ce qui pose un problème majeur d'intégration pour associer les deux fonctionnalités dans un même capteur.

Aussi, nous sommes amenés à discuter des différentes alternatives envisageables, et nous distinguons ci-dessous trois approches possibles que nous avons schématisées sur la Figure IV.4.

A tout imageur CMOS est associé un traitement qui corrige les pixels défectueux, reconstitue les trois couleurs primaires (rouge, vert, bleu), ajuste la balance des blancs, compense les défauts de l'optique (notamment sur les bords), et améliore le rendu visuel. Ces traitements sont accomplis par un processeur compagnon, ou « coprocesseur ».

Ainsi, la première approche d'intégration consiste à ajouter le traitement requis pour l'estimation du mouvement global et plus globalement pour la stabilisation vidéo, au co-processeur (cf Figure IV.4 a.). Cette solution est privilégiée par le secteur industriel pour des raisons de rentabilité de conception.

Une deuxième configuration consiste à déporter une partie de la tâche de stabilisation vidéo au niveau du capteur, soit par traitement numérique dédiée opérant sur le flot de pixels images, soit par traitement au niveau pixel. Le reste de la tâche, notamment les opérations de haut niveau qui requièrent une grande flexibilité de traitement, sont menées par le coprocesseur (cf. Figure IV.4 b.).

Enfin l'essentiel de l'estimation du mouvement global peut être déportée au niveau pixel, comme représenté Figure IV.4 c. Cependant l'opération de stabilisation, pour laquelle la flexibilité des traitements est importante, ne pourra être intégrée qu'au niveau du co-processeur.

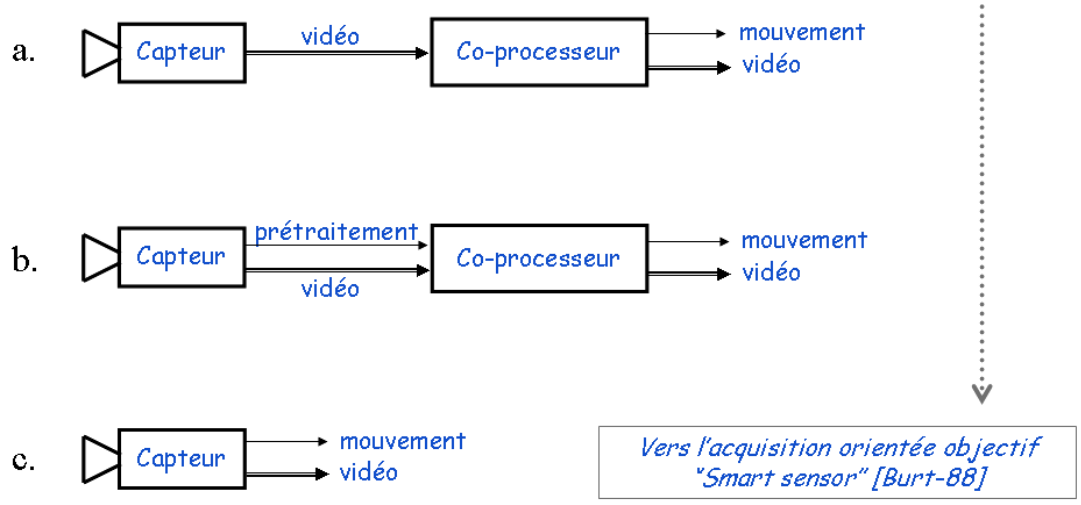


Figure IV.4. Trois approches systèmes possibles : tout le traitement réalisé par le coprocesseur a., report d'une partie du traitement au niveau pixel b., et traitement entièrement réalisé dans le plan focal c.

Dans sa thèse sur l'« évaluation des rétines électroniques pour une définition architecturale d'un système monopuce (SoC) dédié à la vision embarquée », [Elouardi-05] compare quatre types d'architectures systèmes. La première est une rétine à opérateur câblés et un calculateur embarqué, la deuxième est un processeur ARM et une rétine à processeurs analogiques ou/et numériques, la troisième est un processeur ARM associé à un capteur à pixels logarithmiques, et enfin la dernière architecture est un capteur linéaire, un circuit numérique dédié aux prétraitements, un processeur numérique et un système

d'exploitation embarqué. L'auteur montre l'intérêt, pour des applications de filtrage d'image, de la deuxième architecture, donc d'une approche rétine. En effet la rétine utilisée met à profit le parallélisme inhérent à tout capteur d'image en intégrant des traitements élémentaires proches des pixels. Ainsi des traitements de bas niveaux tels que des filtrages, dont la charge de calcul est importante, sont menés en un temps réduit. Par exemple, un filtrage passe-bas sur une image 400×400 nécessite 650 ms à un système constitué d'un imageur et d'un processeur ARM7TDMI, et seulement 40 ms au système rétine et ARM7TDMI.

Par contre, tout ajout « d'intelligence » dans le pixel nécessite d'y ajouter de l'électronique de traitement, donc d'agrandir sa taille. Si l'on considère plus d'un million de pixels, cette surface élémentaire ajoutée est alors multipliée d'autant, ce qui devient évidemment inacceptable. Ainsi il nous faut trouver une architecture qui permette d'allier les avantages d'un traitement niveau pixel et ceux d'un post-traitement.

Comme l'illustre la Figure IV.5 ci-dessous, nous privilégions un report des tâches de détection et de traitement de bas niveau sur le capteur, au niveau pixel, et nous conservons le traitement de haut niveau dans l'architecture digitale associée.

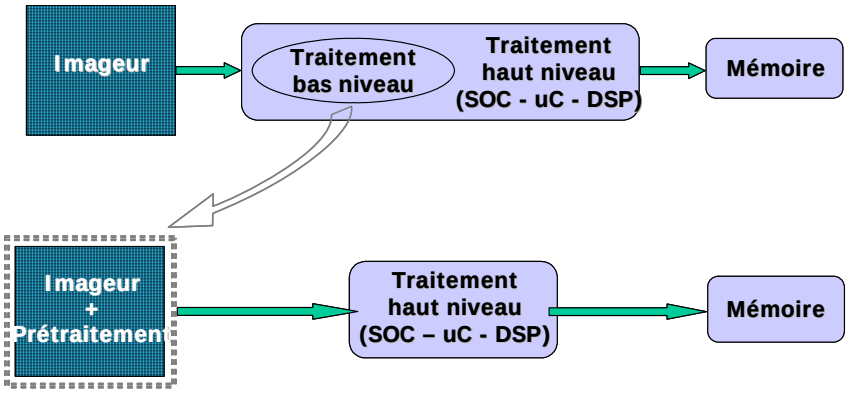


Figure IV.5. Approche privilégiée : report des tâches bas niveau de photodétection « active » sur le capteur.

I.3. Architecture du système sur puce proposé

Comparées aux imageurs, les rétines contiennent des pixels comportant une électronique de prétraitement du signal généralement plus importante et possèdent par conséquent une surface accrue. La résolution du capteur est alors réduite au profit du traitement. Dans le contexte concurrentiel des imageurs pour dispositifs portables que nous avons évoqué au premier chapitre de ce manuscrit, la contrainte de coût minimum est la priorité absolue. Une intégration exclusivement rétine ne peut alors pas être envisagée sur les technologies actuelles CMOS de type « bulk »[80].

[80] *Dans l'avenir, les technologies émergentes d'intégration de type « above IC » lui donnent cependant de nouvelles perspectives.*

Nous proposons de mettre en place une architecture à deux zones photosensibles distinctes : l'une optimisée pour l'image, au centre, et l'autre pour le mouvement, sur la périphérie (cf. Figure IV.6). Nous souhaitons ainsi éviter que les contraintes de conception imposées pour accomplir la fonction d'imagerie n'interfèrent avec celles de l'estimation du mouvement. En effet cela aboutirait à une architecture sous optimale.

La zone photosensible réservée à l'acquisition d'image est alors constituée d'une matrice de pixels de taille minimum et optimisée pour l'application vidéo[81]. Elle est organisée suivant une tessellation carrée.

Les pixels dédiés au mouvement sont, quant à eux, situés sur la périphérie de la zone « image » et sont conçus pour faciliter la perception des mouvements périphériques.

Nous proposons deux techniques pour cela. L'une implémente un nouveau codage de texture basé sur la transformée du recensement décrite au paragraphe II.2. du chapitre précédent. Ces codes sont alors appariés en post-traitement pour obtenir les mouvements locaux. L'autre consiste à extraire le contraste de la scène[82] en continu et à apparier ces contrastes soit en post-traitement, soit au niveau pixel en détectant le passage de ces contrastes d'un pixel à son voisin.

L'avantage essentiel est ici de limiter le post-traitement à réaliser et donc aussi la prise de ressources sur le processeur compagnon. De plus nous associons l'efficacité d'un traitement plan focal à la flexibilité d'un post-traitement digital en distinguant le traitement bas niveau de détection de l'information utile et celui de plus haut niveau de l'interprétation.

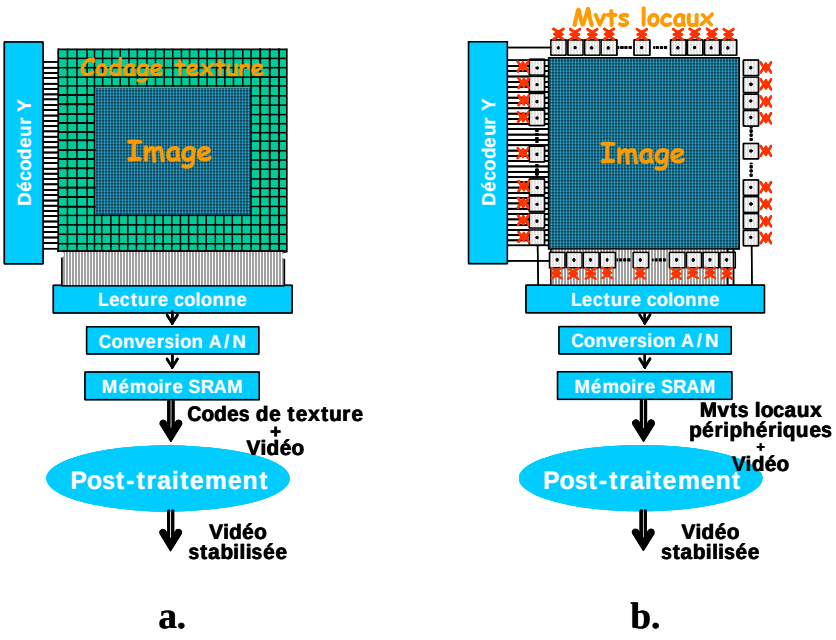


Figure IV.6. Architectures systèmes proposées.

[81] *La taille des pixels est actuellement de l'ordre de 3×3 µm² en lithographie 0.18µm.*

[82] *Information cruciale à toute mesure de mouvement, sans laquelle le mouvement ne peut être perçu (cf Partie I. du Chapitre II.).*

II. VERS L'INTEGRATION DE TRAITEMENTS PERIPHERIQUES DANS LE PLAN FOCAL

En rapport avec les deux approches que nous proposons pour réaliser la mesure des mouvements locaux périphériques, nous nous intéressons ici à l'intégration de la transformée du recensement et à l'extraction de contraste dans le plan focal. Leur mise en correspondance, menée en post-traitement et au niveau pixel, sera considérée dans les sections III. et IV.

Lors d'un traitement d'images effectué au niveau pixel, le bruit présent n'est pas corrigé parce que nous devons minimiser la surface du pixel. Aussi, nous le considérons et étudions son influence sur les performances de nos estimations des mouvements.

Pour cela nous développons un modèle de ce bruit, que nous introduisons ensuite dans nos séquences vidéo de test, ce qui nous permettra alors de quantifier les performances.

II.1. Modélisation du bruit spatial fixe

Nous considérons ici l'architecture d'un pixel APS-3T, que nous étudions par simulation électrique et déterminons un modèle du bruit fixe qui caractérise ce pixel[83].

En effet, les transistors fabriqués dans un procédé CMOS voient leur tension de seuil et leur conductance varier d'un composant à un autre sur le même circuit même s'ils sont exactement identiques lors de la conception. Il en résulte un bruit spatial fixe qui se matérialise par une différence de potentiel observée entre deux pixels d'un imageur CMOS, alors que ces derniers reçoivent exactement la même quantité de lumière.

Dans un pixel APS-3T, le composant principal à l'origine de ce bruit est le transistor suiveur [Degerli-00] [Navarro-03]. Ce sont notamment les dispersions de sa tension de seuil Vt et de son facteur K de transconductance gm qui agissent directement sur le Vgs du transistor et donc sur la tension pixel :

Eq. IV.3.

$$V_{gs} = V_T + \sqrt{\frac{2.I_D}{K.\frac{W}{L}}}$$

Celles-ci se manifestent par un bruit fixe d'un pixel à un autre.

Or la transconductance gm dépend aussi du courant de polarisation du transistor, une dispersion de ce courant entraîne donc aussi une dispersion du Vgs. Ce point de polarisation dépend du courant généré par le transistor de polarisation de colonne qui est lui-même dispersif (« polarisation » sur Figure IV.7). Le courant de polarisation étant commun à chacun des pixels d'une même colonne, il conduit donc à la présence d'un bruit fixe de colonne.

Nous modélisons par conséquent ce bruit fixe sur chacun des pixels par deux composantes « BSF1 » et « BSF2 », qui décrivent respectivement les dispersions de pixel à pixel, et de colonne à colonne.

---

[83] *La structure de mesure de luminance est prise ici à titre d'exemple, le codage du recensement présenté ci-dessous s'intégrant de la même manière sur les structures APS-4T.*

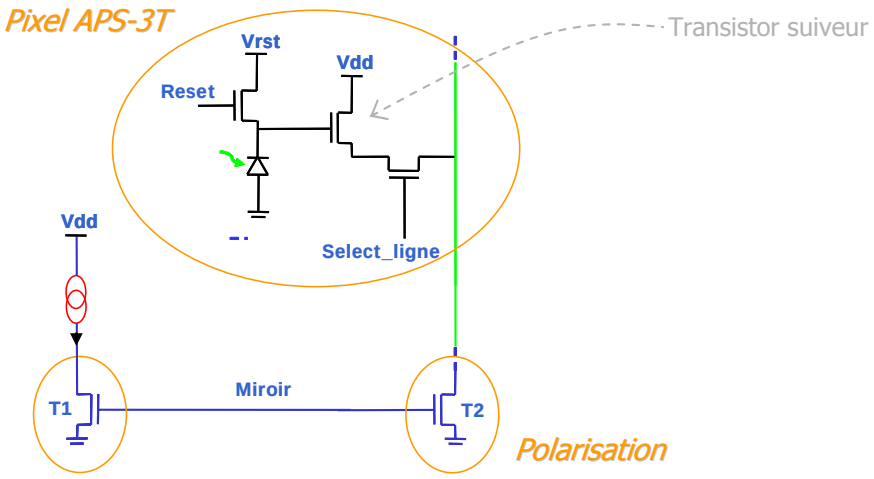


Figure IV.7. Structure de pixel APS-3T simulé, avec polarisation associée commune à chaque colonne de pixels.

Afin que ces constantes soient les plus réalistes possibles, nous avons étudié le pixel en intégrant les cartes modèles de dispersions fournies par le fondeur des circuits intégrés[84]. Les simulations, lancées sous environnement Cadence, sont des études statistiques de type « monte carlo ».

Les tailles des transistors utilisées pour toutes les simulations du pixel sont les suivantes :

$$\frac{W_{reset}}{L_{reset}} = \frac{W_{select}}{L_{select}} = \frac{0.5\ \mu m}{0.35\ \mu m} \qquad \frac{W_{suiveur}}{L_{suiveur}} = \frac{1.6\ \mu m}{0.7\ \mu m} \qquad \frac{W_{T1\_miroir}}{L_{T1\_miroir}} = \frac{W_{T2\_miroir}}{L_{T2\_miroir}} = \frac{10\ \mu m}{1\ \mu m}$$

Les dispersions sur le Vgs du transistor suiveur se déduisent des trois figures ci-après.

La Figure IV.8 présente l'histogramme de la tension Vgs du transistor suiveur du pixel, celui-ci étant polarisé par un générateur de courant idéal, de valeur 1.2 µA. D'après cette figure, nous choisissons de modéliser la composante « BF1 » du bruit fixe (de pixel à pixel) par une Gaussienne de moyenne nulle et d'écart type 8.9 mV.

Nous reportons ensuite sur la Figure IV.9 l'histogramme de dispersion de la copie du courant par un miroir dont la branche de sortie, qui est reliée à la sortie du pixel APS-3T, recopie un courant de 1.2 µA généré par une source idéale. Nous choisissons là aussi de modéliser cette répartition par une Gaussienne de moyenne nulle et d'écart type 94 nA. Puis, à partir de la Figure IV.10, nous remarquons que l'évolution de la tension Vgs du transistor suiveur est quasi-linéaire pour des courants de polarisation de moyenne 1.2 µA et d'écart type 100 nA correspondant aux dispersions de courant de la Figure IV.9. Cela signifie que finalement, nous pouvons modéliser l'influence de la dispersion du courant de polarisation de colonne par une Gaussienne de moyenne nulle et d'écart type 2.5 mV

[84] *Technologie 0.35 µm de chez AMS, à 4 niveaux de métaux et 2 niveaux de polysilicium.*

L'excursion de la tension de sortie du pixel étant voisine de 1.2 V, les deux composantes Gaussiennes « BF1 » et « BF2 » de notre modèle du bruit fixe ont pour écarts types respectifs 0.8 % et 0.2 %.

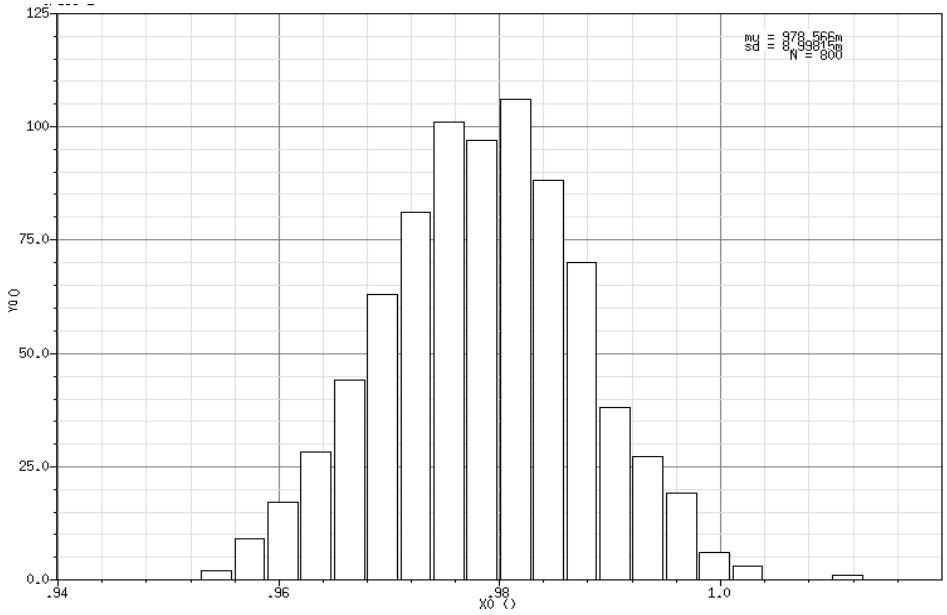


Figure IV.8. Histogramme de la tension Vgs du transistor suiveur de taille W=1.6µm et L=0.7µm sans dispersion de son courant de polarisation (source de courant idéale).

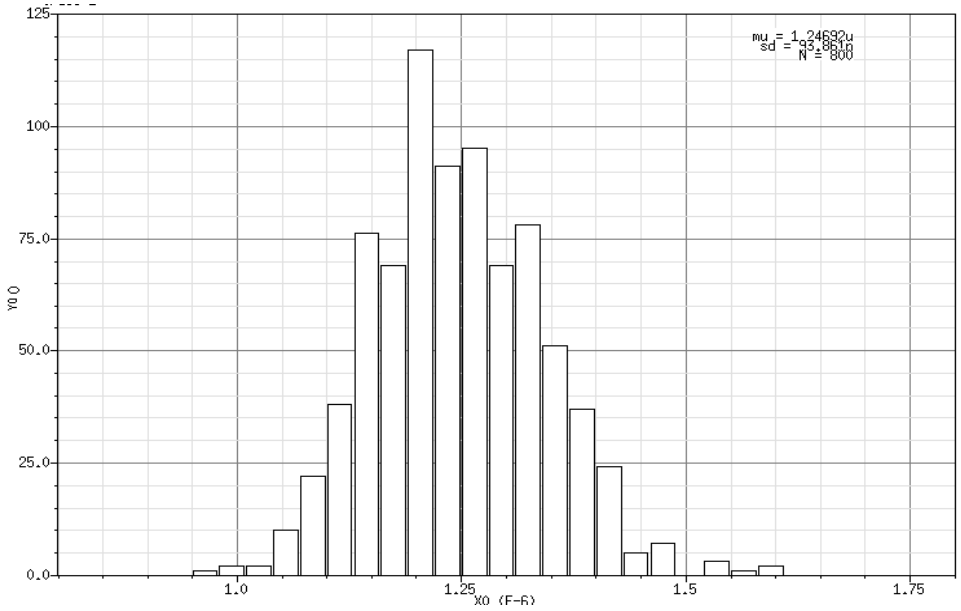


Figure IV.9. Histogramme du courant de polarisation de colonne pour un miroir de courant de taille W=10µm et L=1µm.

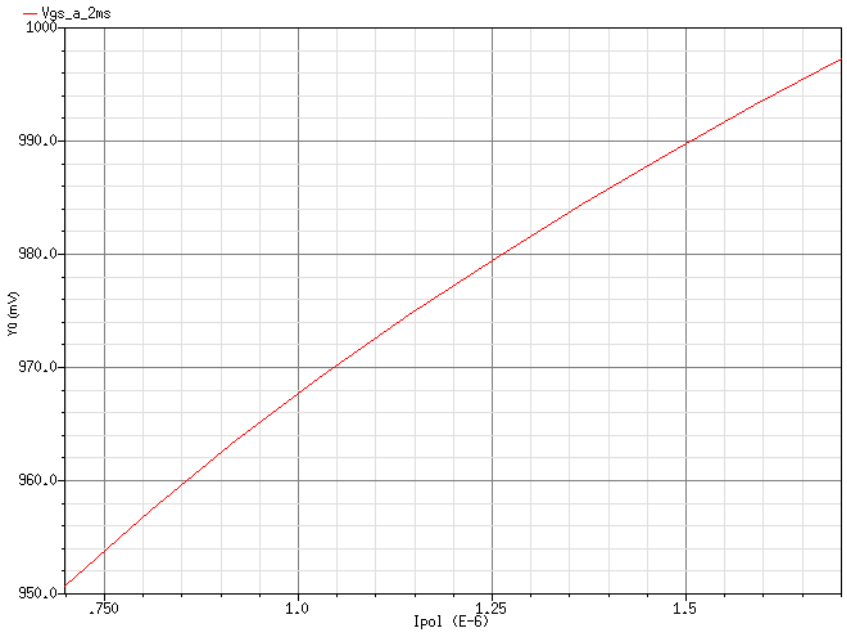


Figure IV.10. Evolution de la tension Vgs du transistor suiveur de taille W=1.6µm et L=0.7µm de l'APS-3T en fonction du courant de polarisation de colonne, autour de 1.2µA.

En pratique, nous ajoutons ce bruit aux valeurs de luminance des pixels appartenant aux séquences réelles et synthétiques paramétrées présentées au chapitre précédent. La première constante ajoutée « BSF1 » est propre à chacun des pixels images, et la deuxième « BSF2 » est propre à chacune des colonnes. Ces deux constantes sont obtenues par un tirage aléatoire unique respectant les deux répartitions Gaussiennes obtenues précédemment.

Nous avons alors bâti des séquences vidéo bruitées qui nous permettent de mieux caractériser les performances du notre technique d'estimation du mouvement global. L'aspect des séquences résultantes est illustré sur la figure suivante, avec et sans réduction du bruit spatial fixe.

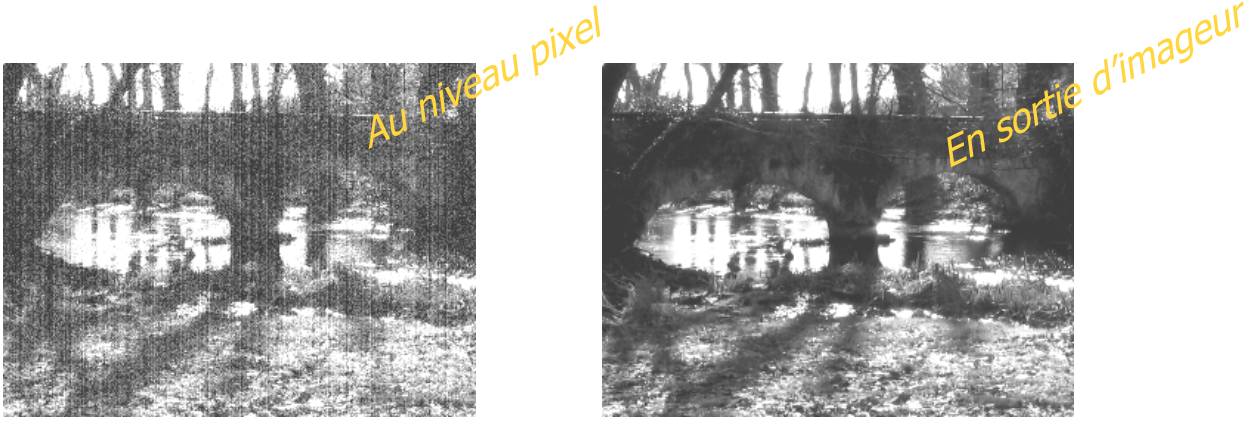


Figure IV.11. Aspect des données à traiter au niveau pixel (à gauche), dans le plan focal, et celles en sortie de l'imageur (à droite).

II.2. Performances obtenues sur données bruitées et améliorations

II.2.a. Transformée du « recensement »

Nous avons étudié et mis en évidence au chapitre précédent les propriétés de la transformée du recensement issue de [Zabih & Woodfill-94], que nous nous proposons maintenant d'intégrer sur silicium. Nous avons relevé au chapitre précédent que la transformée du recensement sur des blocs de taille 3x3 pixels était moins performante que celle de blocs 5×5. Cependant, nous choisissons de réaliser tout d'abord une intégration 3×3 afin de limiter la complexité des pixels lors de cette première étape de prototypage. Cette intégration silicium permettra de réduire la charge de calcul nécessaire à l'estimation du mouvement de 32%[85] et de comparer les performances de cette technique intégrée au niveau pixel avec celles d'une implémentation en post-traitement.

La technique de codage de texture proposée consiste à comparer la luminance d'un pixel par rapport à celles de ses voisins directs. Cependant, la dispersion des caractéristiques des composants électroniques associée à toute fabrication de circuit fait que, dans le cas présent, si deux pixels reçoivent exactement la même quantité de lumière, ils ne restitueront pourtant pas la même information électrique. Ainsi une comparaison codée « 1 » dans une image peut devenir « 0 » dans l'image suivante uniquement à cause du bruit. Nous avons constaté ceci lors des travaux de [Navarro-03], au cours desquels le taux d'appariements justes obtenus sur la puce fabriquée était de 50%.

Par simulation à partir des séquences vidéo synthétiques bruitées, le Tableau IV.6 résume les performances obtenues en termes de robustesse des appariements. Nous y exprimons la proportion d'appariements justes rapportée au nombre total d'appariements calculés. Un appariement est dit « juste » lorsque le mouvement local qu'il décrit correspond à celui effectivement appliqué à la séquence vidéo synthétique.

	Séquences non bruitées	Séquences bruitées
Recensement	85 %	40 %

Tableau IV.6. Taux d'appariements justes par rapport au total, avec ou sans bruit spatial fixe lors des tests.

On s'aperçoit que le taux de bons appariements est de 40% seulement en présence de bruit spatial fixe, ce qui est insuffisant car représente une trop faible densité de mouvements locaux corrects. Ceci nous a amené à rechercher une nouvelle technique de codage qui nous permette d'améliorer ces résultats, tout en restant intégrable sur silicium, ce qui signifie une complexité limitée.

[85] *Nous donnons ici une indication de la réduction de charge de calcul obtenue dans le cas d'un codage impliquant un voisinage 3×3, et une zone de recherche de ± 10 pixels périphériques autour du pixel à apparier.*

- Amélioration de la robustesse au bruit spatial fixe : le recensement ternaire.

Nous avons proposé une transformée spatiale de l'image à trois états au lieu de deux originellement, nous avons alors appelé cette technique le « codage du recensement ternaire ». Son principe est expliqué sur la Figure IV.12 et consiste à attribuer une confiance plus importante aux zones suffisamment texturées des images pour en percevoir le déplacement.

Nous introduisons pour cela un seuil dans la comparaison de la luminance du pixel central avec ses voisins, et nous distinguons alors trois états :

- « X » si le pixel voisin possède une luminance comparable à celle du pixel central (différence inférieure au seuil),
- « 1 » si la luminance du pixel voisin est suffisamment supérieure à celle du central (différence au-delà du seuil),
- « 0 » si la luminance du pixel voisin est suffisamment inférieure à celle du central (différence au-delà du seuil),

Dans l'exemple de la Figure IV.12, le seuil est fixé à 5 % de la dynamique (soit 12.75 niveaux de gris). Ainsi, si nous considérons le pixel central, de valeur de luminance 129, les pixels voisins dont la luminance est telle que « 129-seuil < Lum_voisin < 129+seuil » se trouvent alors dans la zone à moindre confiance.

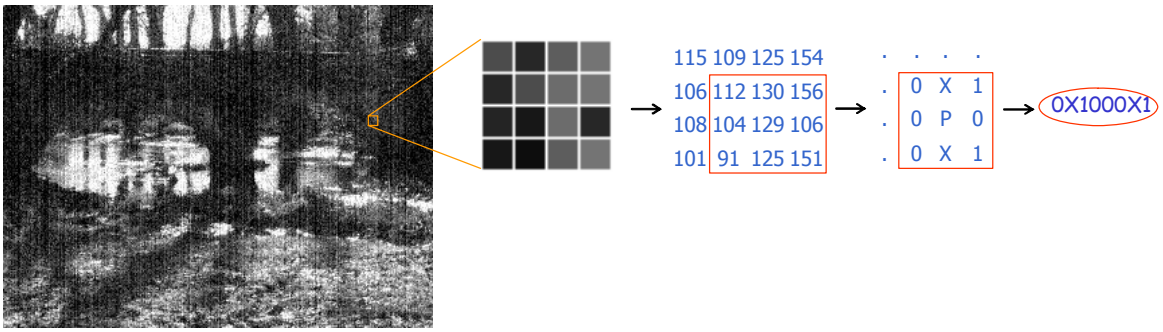


Figure IV.12. Nouvelle technique de codage : le recencement ternaire.

En pratique, nous avons codé ce recensement ternaire en attribuant deux bits par comparaison : « 11 » et « 00 » respectivement pour les états « 1 » et « 0 », et « 10 » pour l'état « X ». Ainsi, la distance de Hamming entre les états « 1 » et « 0 » est deux fois plus importante que celle entre les états « 1 » et « X » ou « 0 » et « X ».

Cette technique, avec un seuil fixé à 5 % de la dynamique pixel, nous permet d'améliorer la robustesse de nos estimations de mouvements locales périphériques de 50%. En effet, le taux d'appariements justes atteint 60% sur séquences bruitées et 82% sur séquences non bruitées (cf. Tableau IV.7).

	Séquences non bruitées	Séquences bruitées
Recensement	85 %	**40 %**
Recensement ternaire	82 %	**60 %**

Tableau IV.7. Taux d'appariements justes par rapport au total, avec ou sans bruit spatial fixe lors des tests, pour le codage du recensement originel et le codage ternaire.

II.2.b. Extraction de contrastes spatiaux par réseaux résistifs

Nous souhaitons réaliser nos mesures des mouvements locaux périphériques par une deuxième technique qui consiste à extraire les contrastes des images, seules zones où la perception du mouvement est possible, pour les mettre en correspondance ensuite. Par manque de temps, nous n'avons pas pu mettre au point une technique efficace de mise en correspondance des contrastes lors de notre étude logicielle. Cette approche n'a donc pu être validée complètement jusqu'ici, mais nous nous attachons à le faire actuellement. Pour autant, nous avons montré à la fin du chapitre précédent l'intérêt d'une détection de contraste dans le plan focal pour la perception du mouvement, c'est pourquoi nous l'étudions ici, légèrement en avance phase par rapport à la validation logicielle.

Nous avons relevé lors de notre état de l'art sur les rétines électroniques deux approches pour détecter les contrastes au niveau pixel. L'une est temporelle et repose sur la détection d'une variation rapide du photocourant pixel. L'autre est spatiale et consiste à détecter une différence non nulle des photosignaux de pixels voisins[86].

L'approche temporelle possède l'inconvénient d'être sensible aux éclairements artificiels tels que les néons, qui se caractérisent par des variations de la lumière émise à la fréquence de 100 Hz [Delbruck & Mead-96]. Ces sources lumineuses sont pourtant couramment employées en scènes intérieures, ce qui nous a conduit à nous intéresser à la technique d'extraction spatiale des contrastes.

Nous choisissons d'extraire ces contrastes en réalisant la différence du photocourant pixel avec un signal image de la moyenne des photocourants des pixels adjacents. Ainsi, si le pixel considéré est soumis à une surexposition (ou une sous-exposition) par rapport aux pixels voisins, la différence du photocourant considéré avec le photocourant moyen sera non nulle. La détection du franchissement d'un seuil par cette différence signalera alors la présence d'un contraste suffisamment perceptible pour être reconnu dans l'image suivante, ou un instant plus tard.

Pour réaliser cette moyenne locale des photocourants, nous employons un réseau résistif tel que celui représenté sur la Figure IV.23 ci-dessous. Le courant $I_e[k]$ est le photocourant du « k^{eme} » pixel de la ligne considérée, et I[k+1] est le courant moyen.

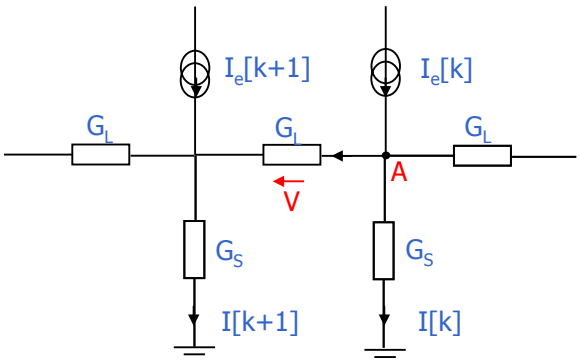


Figure IV.13. Réseau résistif monodimensionnel réalisant le filtrage passe-bas des photocourants.

[86] *En pratique, cette différence devra être supérieure à un certain seuil de détection.*

Nous décrivons lors du paragraphe III.2. l'architecture du pixel que nous avons conçu pour implémenter ce réseau et la différence entre le photocourant du pixel avec le courant moyen, mais le comportement du filtre monodimensionnel ainsi réalisé possède la fonction de transfert suivante :

Eq. IV.4.

$$Hout(z) = \frac{1 + \frac{G_S}{G_L} - (2 + \frac{G_S}{G_L})z + z^2}{1 - (2 + \frac{G_S}{G_L})z + z^2}$$

Nous remarquons qu'il s'agit bien d'un filtre passe-haut, permettant ainsi d'extraire les variations franches de luminosités dans l'image, et dont le comportement dépend du rapport G_S/G_L. Nous verrons plus en détails lors de la section III. suivante que les valeurs de ces admittances permettent d'ajuster le comportement du filtre.

Pour étudier l'influence du bruit spatial fixe sur l'extraction de contraste ainsi réalisée, nous avons là aussi appliqué ce filtre sur des images bruitées. Par contre, à la différence du pixel « recensement ternaire », nous avons choisi de modéliser le bruit du présent au niveau du pixel d'extraction du contraste par des dispersions de pixel à pixel. Nous avons caractérisé par simulation électrique la dispersion de la recopie de courant pour une polarisation à 1.2 nA, en simulation mismatch et suivant le schéma de la Figure IV.14. A partir des histogrammes des différences de courants « I1-I0 » et « I2-I1 » que nous avons obtenus en technologie 0.35 µm de chez AMS (cf. Figure IV.15), nous choisissons de modéliser ce bruit fixe par une dispersion Gaussienne d'écart type de 12% de la valeur nominale et de moyenne nulle.

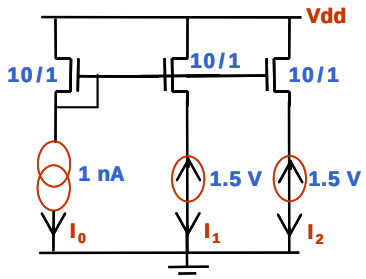


Figure IV.14. Schéma d'étude de la dispersion de courant dans un miroir.

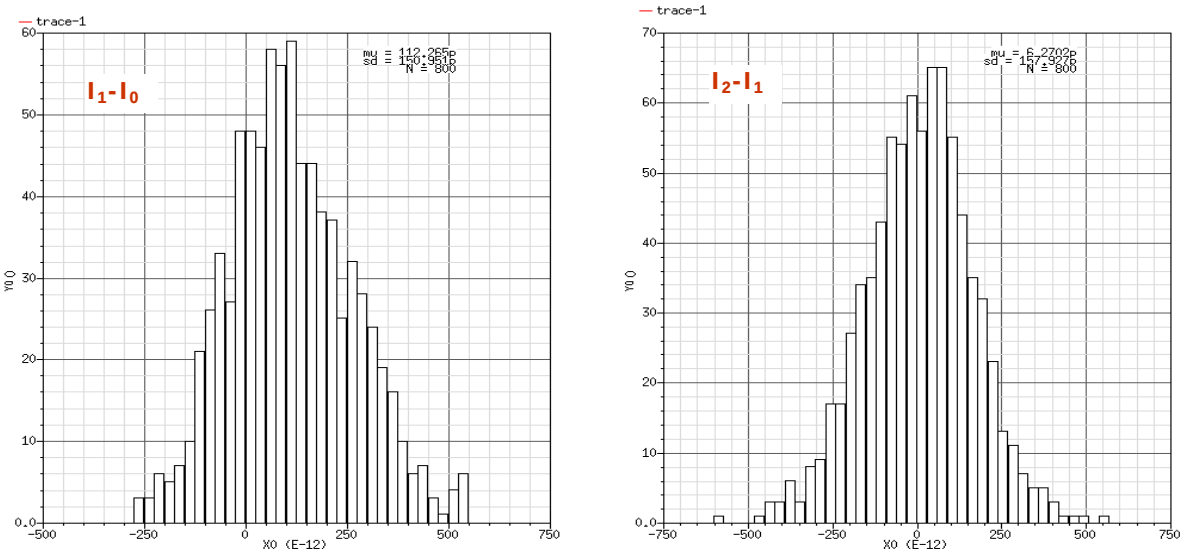


Figure IV.15. Histogramme du courant copié par un miroir de courant de taille W=10µm et L=1µm, pour une valeur nominale de 1.2nA.

Nous présentons sur la Figure IV.16 ci-dessous les résultats d'une extraction de contrastes pour une même image avec le bruit spatial fixe ajouté. Le rapport G_S/G_L est égal à 10.

Figure IV.16. Image bruitée et résultats de l'extraction des contrastes verticaux avec un seuil égal à 3% de la dynamique, et 6% de la dynamique (de gauche à droite).

L'influence principale du bruit spatial fixe sur notre extraction de contraste est de provoquer des détections liées au bruit et non au contenu de la scène.

Pour remédier à ces détections non souhaitées, la solution que nous avons retenue est de tenir compte de l'amplitude de ce bruit dans la valeur du seuil de détection du contraste. C'est-à-dire que nous ajoutons au courant de seuil initial de détection, fixé à 4% de la dynamique, une constante de valeur supérieure à l'écart type de la dispersion Gaussienne du bruit, soit de 11% par exemple. Finalement, notre seuil de détection en courant devra être fixé à 15%

III. INTEGRATION DES TRAITEMENTS PERIPHERIQUES DANS LE PLAN FOCAL

Nous savons, nous l'avons présenté au cours du chapitre I. de ce mémoire, que la technologie CMOS offre la possibilité d'associer sur le même substrat des traitements du signal variés. La photosensibilité du silicium et une électronique de conditionnement spécifique sont mises à profit dans cette partie, au cours de laquelle nous présentons deux architectures de pixels dédiés à la perception du mouvement.

III.1. Codage de texture, transformée du « recensement ternaire »

Pour implanter la variante ternaire de la transformée du recensement que nous avons décrite précédemment, nous avons besoin de comparer la phototension de sortie d'un pixel image de type APS-3T avec celle d'un pixel image voisin en tenant compte d'un « seuil ». Plutôt que d'utiliser un décaleur de niveau associé à un comparateur « standard, » nous avons fait le choix d'utiliser de façon séquentielle un comparateur à hystérésis.

La comparaison avec seuil d'un pixel par rapport à son voisin s'effectue donc en deux temps à l'aide du circuit comparateur comportant sept transistors. Nous présentons la structure du pixel résultant sur la Figure IV.17.

Le seuil est fixé par la largeur du cycle d'hystérésis, qui est réglable d'une part par le courant de polarisation et les dimensions des deux transistors « MN0 » et « MN1 » montés en paire différentielle, et d'autre part par les « amplifications » des miroirs de courant formés par « MP0-MP2 » et « MP3-MP1 »[87].

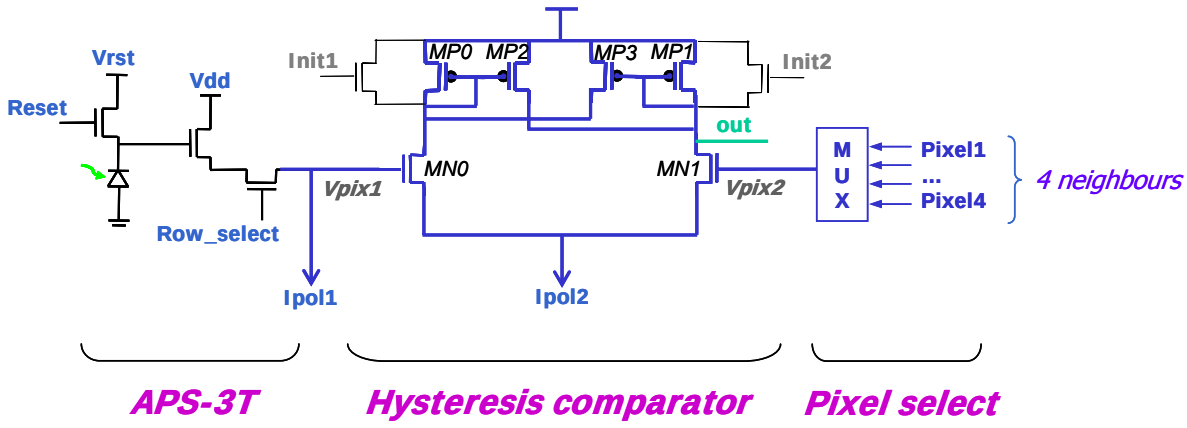
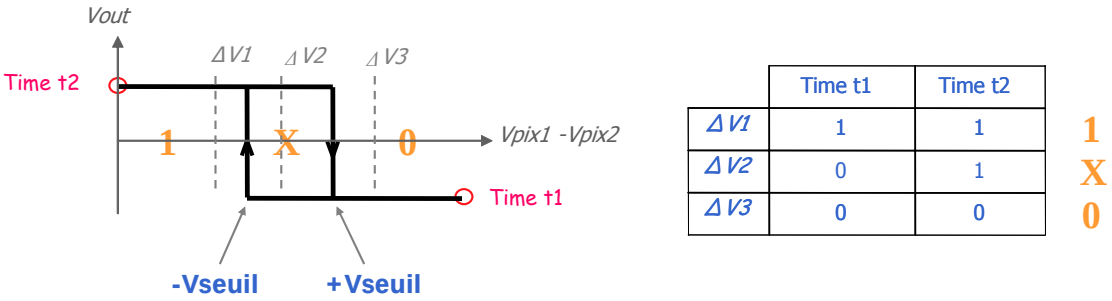


Figure IV.17. Architecture du pixel intégrant la transformée du recensement ternaire.

Le mode opératoire de comparaison est le suivant : nous forçons tout d'abord la branche gauche du comparateur à hystérésis à Vdd par le transistor Init1 (« Time t1 » sur la Figure IV.18), puis la branche est libérée et en observant l'évolution de sa sortie, on pourra ainsi déterminer si la tension du pixel

[87] Deux transistors supplémentaires sont utilisés pour l'initialisation de la structure avant chaque comparaison. Ils ne sont pas représentés ici par souci de clarté.

considéré est inférieure à celle de son voisin moins le seuil (« 1 » sur le cycle d'hystérésis, ΔV1), ou supérieure (« 0 » sur le cycle d'hystérésis, ΔV2 et ΔV3). Le deuxième temps consiste alors à forcer la branche droite du comparateur à hystérésis à Vdd par le transistor Init2 (« Time t2 » sur la Figure IV.18), puis la branche est libérée et la sortie observée, elle vaudra « 0 » si la tension du pixel considéré est supérieure à son voisin plus le seuil (ΔV3), et « 1 » sinon (ΔV1 et ΔV2).



	Time t1	Time t2
ΔV1	1	1
ΔV2	0	1
ΔV3	0	0

1
X
0

Figure IV.18. 2 comparaisons séquentielles pour réaliser le codage du recensement ternaire.

La Figure IV.19 représente les temps de réponses du comparateur, pour une tension de seuil fixée à 115 mV et trois valeurs de tension de mode commun : 1, 1.75, et 2.5. Le temps de réponse tend vers l'infini lorsque l'on se rapproche de la tension de seuil. Puis à partir d'une tension différentielle d'entrée dépassant cette tension de seuil de 3 mV, le temps de réponse est inférieur à 40 ns. Etant donné que nous utilisons le comparateur de façon séquentielle et que nous observons sa sortie après un temps bien précis, nous pouvons définir la tension de seuil en fonction de ce temps de réponse. Par exemple, si nous observons la sortie 40 ns après avoir initialisé le comparateur, et que l'on souhaite mettre en place une tension de seuil à notre codage ternaire de 118 mV, nous définirons la tension de seuil du comparateur à « 118–3=115 mV ».

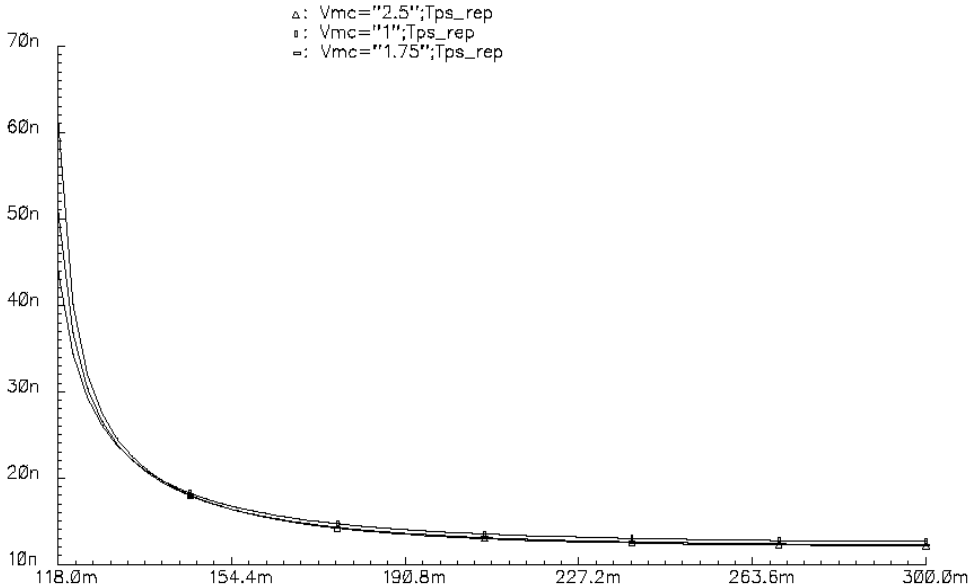


Figure IV.19. Temps de réponse du comparateur en fonction de la tension différentielle d'entrée pour trois modes communs différents : 2.5V, 1.75V et 1V..

Cependant, à cause des dispersions liées au procédé de fabrication, la tension de seuil d'un comparateur à un autre sur le même circuit varie. Ici encore, nous avons donc réalisé une étude statistique

« monte carlo », en mismatch et pour une tension de seuil fixée à 115 mV, dont les résultats sont reportés sur la Figure IV.20. On obtient une tension de seuil moyenne de 116.5 mV, donc une erreur moyenne de 1.5 mV, et un écart type de 8.2 mV.

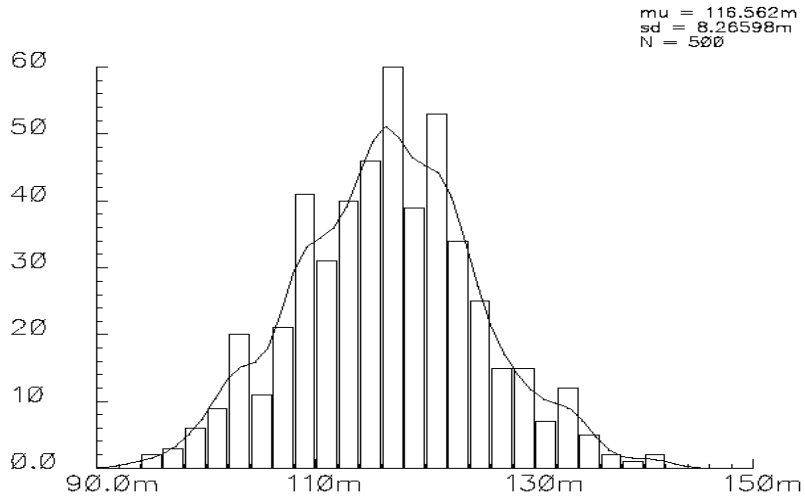


Figure IV.20. Dispersions de la tension de seuil du comparateur.

Cette incertitude sur la valeur de la tension de seuil doit être prise en compte pour fixer la valeur de la tension de seuil de telle manière à nous affranchir aussi de ce bruit.

Nous avons réalisé une estimation de la surface du pixel, elle vaut 25×25 µm² en technologie CMOS 0.35 µm, soit une surface relative de 71×71 λ^2.

III.2. Détection de contrastes orientés

La deuxième architecture que nous proposons est attractive car seules les zones contrastées de notre surface d'intérêt périphérique seront détectées et traitées a posteriori par le coprocesseur. Ce post-traitement est donc optimisé en ne considérant que les données utiles à la perception du mouvement : les contrastes des images.

Sur chacun des côtés de notre zone d'intérêt périphérique, nous souhaitons ne mesurer que les mouvements parallèles à ces bords. Pour cela, nous plaçons sur chaque bord une rangée de pixels « détecteurs de contrastes ». Nous réalisons ces mesures monodimensionnelles en prédisposant le capteur de manière à favoriser la détection des contrastes dont la direction du gradient est parallèle au bord. Ainsi l'amplitude du mouvement local mesuré est effectivement celle perçue au niveau du plan focal. Pour arriver à ces fins, nous mettons en oeuvre des photodétecteurs avec une géométrie spécifique et nous extrayons en temps continu les zones contrastées en interconnectant les photodétecteurs par un réseau résistif implémentant un filtrage spatial passe-haut.

III.2.a. Prédisposition à une détection unidimensionnelle

Un paramètre essentiel qui définit un photodétecteur est sa surface photosensible. Elle définit notamment l'échantillonnage spatial. Il est alors possible de favoriser la sensibilité à certaines formes ou

géométries de la texture présente dans la scène observée. L'élément photosensible réalise alors naturellement et instantanément le calcul. Cette propriété a déjà été exploitée par certains auteurs.

[Barrows & Neely-00] ont notamment développé des capteurs de mouvements monodimensionnels basés sur des photorécepteurs de forme rectangulaire. Ces capteurs de flot optique ont été embarqués sur des engins volants afin de contrôler leur vol et éviter les obstacles. La pertinence de l'emploie de ce type de photorécepteurs monodimensionnels de forme allongée est démontrée dans [Barrows-99]. L'auteur prouve que localement, lorsque l'on met en place une ligne de photodétecteurs[88] dont la longueur « L » est très supérieure à la distance « D » qui les sépare (cf. Figure IV.21), le flot optique perçu est la projection sur l'axe défini par la ligne de photorécepteurs.

Nous adoptons ici une approche similaire et plaçons des photorécepteurs rectangulaires en périphérie des images afin de percevoir les mouvements locaux parallèles aux bords.

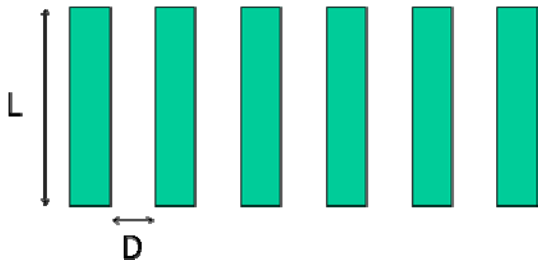


Figure IV.21. Géométrie spécifique des photodétecteurs périphériques.

III.2.b. Extraction des contrastes spatiaux

Nous souhaitons détecter les contrastes de manière analogique et en temps continu, pour les échantillonner en même temps que l'acquisition de l'image. La mesure du mouvement sera ainsi réalisée sur ces données utiles qui ne représentent que quelques pourcents du nombre total des pixels (cf. Chapitre I. section IV.).

La technique que nous avons retenue pour extraire ces fronts de contraste est celle que l'on trouve dans la rétine des vertébrés [Mead-89]. La mesure du contraste en un pixel donné s'obtient en réalisant la différence du photocourant du pixel avec la moyenne locale des photocourants des pixels voisins. Un contraste est présent dès l'instant où la différence est non nulle. Nous illustrons ce procédé sur la Figure IV.22 suivante.

[88] *Au moins une paire de photodétecteurs est nécessaire.*

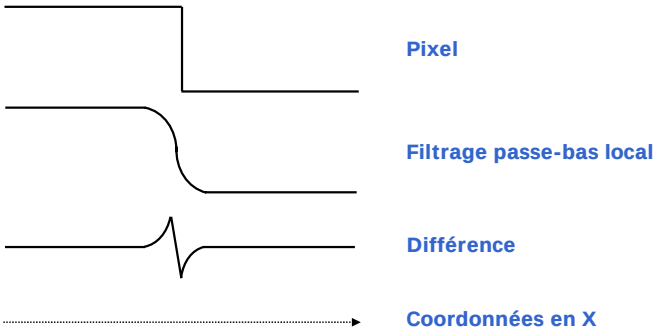


Figure IV.22. Cas monodimensionnel d'extraction des contrastes sur silicium.

Pour réaliser le filtrage passe-bas local des photocourants, ou moyenne locale, nous employons un réseau résistif tel que celui représenté sur la Figure IV.23 ci-dessous, ou chaque conductance sera matérialisée par un transistor MOS en régime linéaire.

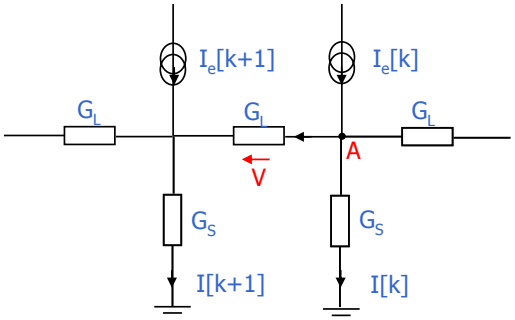


Figure IV.23. Réseau résistif monodimensionnel réalisant le filtrage passe-bas des photocourants.

Notre matrice de pixels réalise un échantillonnage spatial de la scène. Aussi, nous déterminons le comportement du filtre que représente notre réseau comme un système échantillonné.

A partir des deux équations aux différences Eq. IV.5 et Eq. IV.6 suivantes, qui décrivent respectivement la différence de potentiel « V » aux bornes de la conductance G_L au centre du réseau et la somme des courants sur le nœud « A », nous en déduisons la fonction de transfert en z du filtre.

Eq. IV.5. $$I_L[k] = -\frac{G_L}{G_S} \times \left[I[k+1] - I[k] \right]$$

Eq. IV.6. $$I_E[k] + I_L[k-1] = I_L[k] + I[k]$$

D'où : $$-\frac{G_L}{G_S} I[k] + \frac{G_L}{G_S} I[k-1] + I_E[k] = -\frac{G_L}{G_S} I[k+1] + \frac{G_L}{G_S} I[k] + I[k]$$

C'est-à-dire : $$(1 + 2\frac{G_L}{G_S}) I[k] - \frac{G_L}{G_S} I[k-1] - \frac{G_L}{G_S} I[k+1] = I_E[k]$$

En transformée en z : $$I[z] \times \frac{G_L}{G_S} \left[(2 + \frac{G_S}{G_L}) - z^{-1} - z \right] = I_E[z]$$

D'où la fonction de transfert en z du filtre passe-bas :

Eq. IV.7. $$H(z)=\frac{I[z]}{I_E[k]}=\frac{-\frac{G_S}{G_L}}{1-(2+\frac{G_S}{G_L})z+z^2}$$

Enfin, nous avons défini notre détection du contraste comme résultant de la différence « I - I_E ». Ainsi, nous avons :

$$Iout=I_E-I=I_E-H.I_E=I_E(1-H)$$

La fonction de transfert décrivant l'extraction de contraste s'écrit donc :

$$Hout(z)=\frac{Iout(z)}{I_E(z)}=1-H(z)=1-\frac{-\frac{G_S}{G_L}}{1-(2+\frac{G_S}{G_L})z+z^2}$$

C'est à dire sous la forme de l'équation Eq. IV.8 : $$\boldsymbol{Hout(z)=\frac{I[z]}{I_E[k]}=\frac{1+\frac{G_S}{G_L}-(2+\frac{G_S}{G_L})z+z^2}{1-(2+\frac{G_S}{G_L})z+z^2}}$$

Nous représentons sur la Figure IV.24 ci-après le schéma électrique simplifié du pixel réalisant cette extraction du contraste.

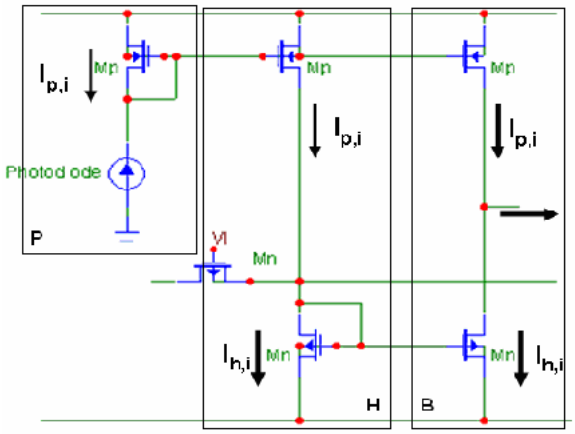


Figure IV.24 : Schéma électrique du pixel avec filtrage spatial passe-haut monodimensionnel.

Le bloc « P » effectue la conversion lumière - courant (photodiode) et recopie son courant en entrée des blocs « H » et « B » via un miroir de courant. Le bloc « H » récupère le courant $I_{p,i}$ et le diffuse dans le réseau résistif. Son courant de sortie correspond alors à une moyenne des courants circulant dans les pixels adjacents. Le courant $I_{h,i}$ issu de la photodiode et de la diffusion des courants circulant dans les pixels adjacents est alors recopié en entrée du bloc « B ». Le bloc « B » effectue alors la différence entre

ces deux courants et signale la présence d'un front de contraste grâce à la variation de tension sur le nœud de sortie.

Nous étudions actuellement le comportement électrique de la structure, en termes de bruit et de temps de réponse, afin de valider ou pas sa structure pour la détection de contraste. Nous estimons la surface du pixel à 22×22 µm² en technologie CMOS 0.35 µm, soit 62×62 λ².

Déterminons maintenant la charge de calcul nécessaire pour mettre en correspondance ces lignes de pixels avec extraction de contraste afin de mieux cerner le post-traitement qu'il nous faudra mettre en place pour estimer les mouvements locaux périphériques.

La sortie de chaque pixel est un bit qui est à « 0 » ou « 1 » selon qu'un contraste soit détecté ou pas. Ainsi comme nous l'illustrons sur la Figure IV.25, la mise en correspondance d'un bloc de « M » pixels met en jeu deux opérations de « M » bits, soit deux opérations.

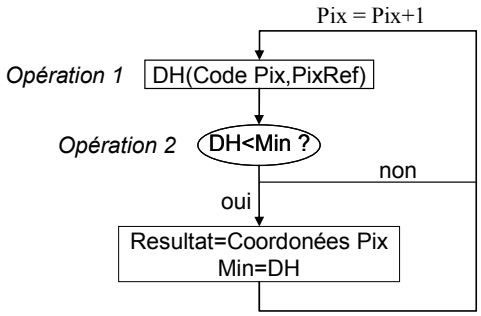


Figure IV.25. Mise en correspondance des blocs de pixels avec extraction de contraste.

Avec une première approche qui consiste à déterminer en une dimension, et sur une zone de recherche de taille « 2.S+1 » pixels, le minimum de dissemblance entre blocs de pixels d'une ligne (ou colonne), le calcul de « N » mouvements locaux périphériques nécessite « **2.(2.S+1).N » opérations élémentaires.** Cela est très modeste comparé à la technique du « block matching » ou à la mise en correspondance de transformée du recensement que nous avons décrites précédemment (recherche en deux dimensions).

IV. ADÉQUATION ALGORITHME ARCHITECTURE

Nous réalisons dans cette section une synthèse des techniques et architectures d'estimation du mouvement global que nous avons étudiées. Pour chacune d'elle nous évaluons de manière qualitative l'adéquation entre l'algorithme à implémenter et l'architecture retenue, en termes de surface ajoutée et de charge de calcul restant à accomplir en post-traitement.

Rappelons également il est nécessaire de prévoir une surface supplémentaire sur la matrice de pixels « image » pour permettre le recadrage de la vidéo : elle s'élève à 12% de la surface d'acquisition requise sans la fonction de stabilisation vidéo, et dépend de l'amplitude du mouvement de « bougé » à compenser. Nous ne tenons pas compte dans notre analyse d'adéquation algorithme architecture qui suit de cette surface qui est commune à toute technique de stabilisation vidéo électronique.

En revanche, à celle-ci s'ajoute la surface qu'occupent les pixels périphériques dédiés à la perception du mouvement. Nous reportons dans le Tableau IV.8 ce coût en surface ajoutée, proportionnellement à la surface de la matrice de pixels dépourvue de cette zone périphérique, ainsi que la charge de calcul restante pour finaliser l'estimation du mouvement global, en fonction du nombre de mouvements locaux et des paramètres associés à leur technique d'estimation.

Nous avons regroupé dans le tableau les trois solutions que nous avons étudiées jusqu'ici, ainsi qu'une autre qui nous semble intéressante. Elle exploite les détecteurs du mouvement développés par [Kramer-96], et plus précisément le type « FTI » (cf. Chapitre II.) pour fournir les données des mouvements locaux périphériques à partir desquelles le mouvement global est extrait.

Technique	**Précision sur l'EMG**		**Surface périphérique en % de la surface initiale (Technologie λ µm)**	**Post-traitement (op.)**
	Moy.	*Ecart type*		
Block-matching 2D	1 %	25 %	0	$8.M^2.N.S^2 + 42.N + 207$
Recensement ternaire 2D + post-traitement	12 %[89]	8 %	$\frac{1}{4}$	$N.(2.S+1)^2 + 42.N + 207$
Contrastes 1D + post-traitement	-	-	$\frac{1}{L.l}.\left[62.\lambda.(L+l) + 3844.\lambda^2\right]$	$2.(2.S+1).N + 42.N + 207$
Contrastes 1D + traitement pixel + post-traitement	-	-	$\frac{1}{L.l}.\left[83.\lambda.(L+l) + 6889.\lambda^2\right]$	$42.N + 207$

Tableau IV.8. Synthèse des techniques des estimations du mouvement global inter images proposées. « L.l » est la surface de la matrice de pixels « image », « S » l'amplitude de la zone de recherche, « M » la taille des blocs de pixels ou de codage, et « N » le nombre de mouvements locaux périphériques considérés.

Dans le cas des solutions 1D (les deux techniques du bas du tableau), nous obtenons l'expression de la surface périphérique, en fonction de la surface occupée par la matrice de pixels sans cette zone périphérique (surface initiale qui vaut « L.l »), en considérant la taille constante de la zone périphérique ajoutée, qui est liée à la taille d'un pixel (62 λ et 83 λ). Soit « T » la taille d'un pixel périphérique, la surface totale vaut alors :

Surface totale = (L+T).(l+T) = L.l + T.(L+l)+T²

La surface périphérique est quant à elle égale à :

Surface périph. = Surface totale – Surface initiale = T.(L+l)+T²

Et enfin le rapport de la surface de la zone périphérique par la surface initiale vaut :

[89] *Cette précision est donnée relativement à la taille pixel. Etant donné qu'un pixel ternaire qui est carré, possède un côté de taille quatre fois plus grande qu'un pixel image, la précision à considérer en pratique pour la stabilisation d'une séquence d'images doit donc être multipliée par quatre, soit de 48%.*

$$\frac{Surface\ périphérique}{Surface\ totale} = \frac{1}{L.l}.\left[T.(\mathrm{L}+\mathrm{l})+T^2\right]$$

Par contre, la surface de la zone périphérique dédiée à la transformée du recensement est proportionnelle à la surface initiale. En effet, pour être capable de déterminer des mouvements d'amplitudes allant jusqu'à 3% de la taille image, la largeur de la bande de pixels périphériques doit donc mesurer « 2.3 = 6 % » de la longueur (ou de la largeur) de l'image. Ainsi l'équation III.3. que nous avons développée au chapitre précédent nous donne une proportion de 25 % de la surface initiale.

Prenons l'exemple d'une séquence vidéo SVGA (800×600), avec une amplitude de mouvement de 20 pixels (3% de la taille image), un nombre N = 280 mesures de mouvements locaux périphériques et une taille de blocs de M×M pixels, nous obtenons les caractéristiques ci-dessous :

Technique	**Précision sur l'EMG**		**Surface périphérique en % de la surface initiale (Technologie 0.13 µm)**	**Post-traitement (op.)**
	Moy.	*Ecart type*		
Block-matching 2D	1 %	25 %	0 %	94 147 967
Recensement ternaire 2D + post-traitement	12 %[90]	8 %	25 %	482 647
Contrastes 1D + post-traitement	-	-	0.8 %	34 927
Contrastes 1D + traitement pixel + post-traitement	-	-	1 %	11 967

Tableau IV.9. Synthèse des techniques des estimations du mouvement global inter images proposées en technologie CMOS 0.13 µm, et dans le cas d' une séquence vidéo 800×600, d'une amplitude de mouvement de 3% de la taille image (20 pixels), d'une taille de bloc de 5 pixels, et un nombre N =280 mouvements locaux.

Nous sommes capables à la lecture des tableaux IV.4. et IV.5, du Tableau IV.8 et plus particulièrement du Tableau IV.9 ci-dessus, d'avancer quelques conclusions quant aux solutions que nous avons étudiées.

La première technique, de mise en correspondance de blocs (« ou block matching ») est la plus performante avec une précision voisine du pourcent en moyenne. Elle n'engendre qui plus est pas de surcoût de surface périphérique, mais le prix à payer se situe au niveau de la charge du post-traitment à mettre en place : près d'une centaine de millions d'opérations élémentaires à réaliser pendant un temps inter trame de 33 ms. Cela signifie une puissance de calcul proche de 3 milliards d'opérations élémentaires par seconde !

[90] *Cette précision est donnée relativement à la taille pixel. Etant donné qu'un pixel ternaire qui est carré, possède un côté de taille quatre fois plus grande qu'un pixel image, la précision à considérer en pratique pour la stabilisation d'une séquence d'images doit donc être multipliée par quatre, soit de 48%.*

La deuxième technique du Tableau IV.9, qui consiste à intégrer la transformée du recensement ternaire au niveau pixel, permet de réduire la charge de calcul nécessaire pour estimer le mouvement global inter images de plus de 50 %. Ce résultat est au prix d'une surface périphérique additionnelle équivalente au quart de la surface initiale de la matrice de pixels, ce qui est rédhibitoire quant à son choix. Ses performances logicielles sont qui plus est trop faibles.

En ce qui concerne les deux solutions basées sur une perception du mouvement unidimensionnel, elles s'avèrent très attractives en termes de charge de calcul puisqu'elles requièrent des ressources matérielles capables d'accomplir entre 350 mille opérations et 1 million d'opérations par seconde. Leur surface ajoutée est somme toute très raisonnable : voisine de 1%, ce qui est tout à fait acceptable pour un fabriquant d'imageurs. Les performances n'ont toutefois pas pu être validées de façon logicielle pour l'heure, ce qui représente une grande interrogation. Un autre aspect qui demande à être validé concerne la quatrième solution, il s'agit de la synchronisation des mesures locales du mouvement à l'aide des détecteurs de [Kramer-96].

Finalement, nous avons proposé différentes solutions d'intégration du système d'estimation du mouvement global inter images, et le choix final entre les solutions 1, 3, ou 4 résultera de compromis qui doivent être trouvés entre performances, surface, et ressources matérielles requises.

CONCLUSION

Nous avons présenté dans ce dernier chapitre la dernière phase de notre démarche de conception d'un capteur d'image dédié à l'acquisition vidéo et la mesure du mouvement global inter images en vue le la stabilisation vidéo.

Ceci a consisté tout d'abord en la définition de l'architecture générale du système, c'est-à-dire le partitionnement de la chaîne de traitement du signal que nous avons validé de façon purement algorithmique au chapitre précédent.

Pour cela nous avons étudié la charge de calcul requise pour une intégration sur un processeur. Celles-ci étant définies, nous avons proposé une partition décomposée en un traitement au niveau pixel dédié à la mesure des mouvements locaux périphériques, très coûteux en calculs, et un post-traitement pour réaliser notamment la tâche d'estimation du mouvement global et le calcul de la correction du mouvement global à appliquer par recadrage à la vidéo pour la stabiliser.

Nous avons ensuite considéré l'intégration silicium des traitements au niveau pixel et étudié l'influence du bruit présent à ce niveau de la chaîne de traitement du signal sur les performances de nos techniques d'estimation des mouvements locaux. Cela nous a conduit à améliorer la robustesse des techniques initiales, avant de poursuivre par la conception de ces pixels dédiés à la perception du mouvement, encore appelés « détecteurs locaux du mouvement ».

Enfin nous avons évalué les différentes solutions proposées d'un point de vue adéquation algorithme architecture, et avons mis en exergue les compromis à trouver pour choisir l'architecture finale.

CONCLUSION GENERALE

Les capteurs d'images ont connu au cours de la dernière décennie une croissance exceptionnelle, qui est essentiellement liée à l'avènement des applications multimédia et des dispositifs portables embarquant ces systèmes (téléphones mobiles et « PDA » par exemples). La technologie de fabrication CMOS, de par sa faible consommation et sa capacité d'intégration, connaît un franc succès face au CCD qui reste, certainement pour peu de temps, la technologie reine de l'imagerie. Les fabricants de semi-conducteurs, comme STMicroelectronics, ont même développé des procédés de fabrication CMOS dédiés à l'imagerie, avec des propriétés optiques spécifiques. Cependant, les difficultés et les coûts mis jeu pour les mettre au point et concevoir des capteurs qui réalisent une acquisition d'image de qualité sont énormes. Il s'agit pourtant d'une fonction que l'on pourrait considérer maîtrisée, mais les contraintes d'encombrement sont très fortes. L'association de ces contraintes avec l'augmentation continue du nombre de pixels sur un même imageur, qui dépasse aujourd'hui le million d'éléments, fait qu'il est aujourd'hui impossible, a priori, d'intégrer une nouvelle fonction de traitement d'image ou de vision autrement qu'en post-traitement des images.

Par ailleurs, le marché des capteurs d'images destinés à être embarqués dans les dispositifs portables tels que les téléphones ou les « PDA » est à la fois porteur et concurrentiel. Aussi, l'ajout de nouvelles fonctions constitue un argument de vente supplémentaire. C'est dans ce contexte que STMicroelectronics s'intéresse à la stabilisation vidéo, et que nous avons étudié des solutions pour intégrer cette fonctionnalité à un imageur.

Nous avons commencé notre travail par un état de l'art sur les capteurs d'images et les rétines CMOS afin de mieux appréhender les différentes contraintes de conception. Nous avons notamment présenté les architectures de conditionnement de l'information photoélectrique. Outre le fait qu'elles sont liées à l'application visée, il en est ressorti que la structure des capteurs d'images, constituée de cellules élémentaires parallélisées que sont les pixels, est bien adaptée aux traitements d'images de bas niveau.

Nous nous sommes intéressé ensuite à la tâche de perception et d'estimation du mouvement, requise pour mener à bien la stabilisation vidéo par traitement d'images. Après un état de l'art sur les techniques existantes de stabilisation vidéo ainsi que sur les capteurs du mouvement développés jusqu'ici, nous avons pu tirer des conclusions quant au choix de la technique d'estimation du mouvement. Nous avons souligné en particulier les bonnes performances des techniques de mise en correspondance.

A partir de là, une troisième étape a consisté à formaliser les tâches de perception du mouvement et de stabilisation vidéo électronique. Après avoir défini les spécifications et les contraintes matérielles de notre système, nous avons proposé une technique d'estimation du mouvement global inter images. Cette technique considère un modèle de mouvement global inter images à quatre paramètres et consiste à extraire le mouvement global à partir de mesures locales du mouvement en périphérie de la zone d'acquisition d'image. Cette solution s'est avérée originale et performante. Nous l'avons en effet validée dans un premier temps de manière logicielle, en adoptant un plan d'expérience comprenant des séquences vidéo synthétiques et réelles. Elle atteint ainsi une précision d'estimation du mouvement global de 1%, et nous permet d'obtenir une erreur absolue sur le mouvement à compenser toujours inférieure au pixel. Ces résultats sont compatibles avec l'application visée de stabilisation vidéo.

Nous avons ensuite envisagé l'intégration matérielle de cette technique à un imageur. Nous avons donc étudié l'adéquation de l'algorithme à l'architecture, en proposant un partitionnement des traitements

qui permette d'associer les avantages d'un traitement au niveau pixel avec ceux d'un post-traitement. Avant de concevoir les pixels périphériques à la zone d'acquisition d'image, dédiés à la perception du mouvement, nous avons caractérisé les performances de nos techniques d'estimation des mouvements locaux en présence de bruit. En effet au niveau pixel, il faut considérer les défauts (bruit spatial fixe, bruit de reset, par exemple) qui sont normalement corrigés en sortie de l'imageur et ne sont donc pas pris en compte dans les algorithmes de traitement d'images. Ceci nous a alors amené à apporter des modifications aux techniques que nous avions initialement choisies et ce afin d'augmenter leurs robustesses vis à vis de ces défauts.

Enfin, nous avons présenté une synthèse des trois architectures de système que nous proposons, plus une que nous envisageons. Nous avons alors mis en avant les considérations relatives à l'Adéquation Algorithme Architecture, c'est-à-dire qui concernent la surface, les performances, et les ressources matérielles requises. Cette analyse a montré que le partitionnement que nous avons mis en place semble judicieux, avec une surface périphérique ajoutée faible, de l'ordre de 1 %, et des post-traitements relativement peu contraignants puisqu'ils restent inférieurs au million d'opérations élémentaires par seconde.

Ces résultats sont les fruits d'une approche originale de conception qui considère le capteur comme un système jouissant de capacités de traitement variées pouvant être associées pour réaliser un traitement optimal pour l'application visée. Le capteur « intelligent » qui en résulte est une « rétine ». Dans notre cas, nous avons proposé d'associer sur le même substrat deux types de pixels, l'un spécialement conçu pour l'acquisition d'image, et l'autre pour le mouvement.

En abordant un problème de cette manière, une modélisation du système, même partielle, s'avère nécessaire afin de formaliser et valider les concepts. Celle-ci vient en amont dans le flot de conception du capteur d'image, qui comprend ensuite le développement des algorithmes de traitement et leur intégration sur le silicium.

En ce qui concerne les perspectives de notre travail, la partie relative à l'intégration silicium est relativement avancée, mais la perspective logique que nous souhaiterions mener à court terme est la réalisation d'un prototype afin de vérifier nos validations logicielles par des mesures sur circuit. Nous devrons pour cela terminer la caractérisation par simulations des pixels de détection des contrastes, et développer un nouvel algorithme d'estimation des mouvements adapté à ces données numérisées.

Par ailleurs, l'une des quatre variantes d'architectures que nous avons décrites à la fin du quatrième chapitre met en œuvre des senseurs locaux de mouvement. Bien qu'il s'agisse d'une architecture de capteur déjà évoquée et étudiée par Jorg Kramer lors de ses travaux à Caltech dans les années 1990, la principale difficulté réside dans la synchronisation (ou plus exactement l'intégration) des informations issues de ces capteurs, par nature asynchrones, avec la fréquence trame. Il s'agit ici d'une perspective plutôt à moyen terme pour laquelle une validation préliminaire des architectures de pixels devra être réalisée.

Une dernière possibilité peut s'avérer intéressante et être mise en œuvre rapidement dès l'instant que le capteur embarque de la compression vidéo. En effet, les traitements mis en œuvre lors d'un codage MPEG par exemple réalisent une estimation du mouvement de blocs de pixels. Il serait donc

intéressant de réutiliser ces informations pour en extraire le mouvement global inter trames à l'aide de notre algorithme d'estimation du mouvement global.

La tendance actuelle est à la course vers des capteurs restituant des images de résolution toujours plus importante, dans un volume toujours plus réduit. Cependant, la taille des pixels développés aujourd'hui, de l'ordre de 2 µm de côté, côtoie dorénavant les limites des composants optiques intégrés, alors que la réduction des dimensions élémentaires des circuits CMOS se poursuit au rythme de la loi de Moore. Ainsi, l'approche « rétine » qui consiste à « sacrifier » de la surface photosensible pour la remplacer par des circuits de traitement placés « au plus près du pixel » commence-t-elle à prendre de l'intérêt aux yeux des industriels, puisque la place à sacrifier l'est déjà du fait des limites optiques. De plus, l'émergence des technologies d'intégration en trois dimensions de type « above IC » ouvrent de nouvelles voies aux capteurs d'images « rétines ».

BIBLIOGRAPHIE

ACIVS-05

➙ *Proceedings of the International IEEE conference on Advanced Concepts for Intelligent Vision Systems, Sept. 20-23, 2005, Antwerp, Belgium.*

Adelson & Bergen-85

➙ *E.H. Adelson and J.R. Bergen, "Spatiotemporal energy models for the perception of motion", Journal of Optical Society of America, vol. 2, n°2, February 1985, pp. 284-299.*

Andersson et al.-02

➙ *K. Andersson, P. Johansson, R. Forchheimer, H. Knutsson, "Backward-forward motion compensated prediction", Advanced Concepts for Intelligent Vision Systems, ACIVS'02, Ghent, Belgium, 2002, pp. 260-267.*

Andreou et al.-91

➙ *A. Andreou, K. Strohbehn, R.E. Jenkins, "Silicon retina for motion computation", IEEE Int. Symposium on Circuits and Systems, Vol. 3, June 1991, pp. 1373-1376.*

Andreou & Boahen-95

➙ *A.G. Andreou, K.A. Boahen, "A 590,000 transistor 48,000 pixel, contrast sensitive, edge enhancing, CMOS imager-silicon retina", Proceedings of Advanced Research in VLSI, March 1995, pp. 225-240.*

Arreguit et al.-96

➙ *X. Arreguit, F.A. Van Schaik, F.V. Bauduin, M. Bidiville, E.A. Raeber, "CMOS motion detector system for pointing devices", IEEE Journal of Solid-State Circuits, vol. 31, n° 12, Dec. 1996, pp. 1916-1921.*

Auberger & Miro-05

➙ *S. Auberger and C. Miro, "Digital video stabilization architecture for low cost devices", Int. Symp. On Image and Signal Processing and Analysis, 15-17 sept. 2005, pp. 474-479.*

Aubert et al.-99

➙ *G. Aubert, R. Deriche and P. Kornprobst, "Computing optical flow via variational techniques", SIAM Journal on Applied Mathematics, vol. 60, n° 1, 1999, pp. 156-182.*

Baker & Matthews-04

➙ *S. Baker and I. Matthews, "Lucas-Kanade 20 years on: a unifying framework",International Journal of Computer Vision, Vol. 56 , n° 3, February-March 2004, pp. 221-255.*

Bandyopadhyay et al.-06

➙ *A. Bandyopadhyay, J. Lee, R.W. Robucci, P. Hasler, "MATIA: A Programmable 80 _W/frame CMOS Block Matrix Transform Imager Architecture", Int. Journal of Solid-State Circuits, Vol. 41, n° 3, March 2006, pp. 663-672.*

Barlow & Levick-65

➙ *H.B. Barlow and W.R. Levick, "The mechanism of directionally sensitive units in rabbit's retina", J. Physiol, No. 178, 1965, pp. 477-504.*

Barron et al.-94

→ *J.L. Barron, D.J. Fleet, S.S. Beauchemin, "Performance of optical flow techniques", Int. Journal of Computer Vision, 1994, vol. 12, pp. 43-77.*

Barrows-99

→ *G.L. Barrows, "Mixed-mode VLSI optic flow sensors for micro air vehicle", PHD thesis, University of Maryland, 1999.*

Barrows & Neely-00

→ *G.L. Barrows and C. Neely, " Mixed-mode VLSI optic flow sensors for in-flight control of a micro air vehicle", SPIE 45th Annual Meeting in San Diego, CA, July 31 - August 4, 2000.*

Benson & Delbruck-91

→ *R.G. Benson and T. Delbrück, "Direction-selective silicon retina that uses null inhibition", Advances in Neural Information Processing Systems 4, D.S. Touretzky, Ed., San Mateo, CA: Morgan Kaufmann, 1991, pp. 756-763.*

Benthien et al.-00

→ *S. Benthien, T. Lule, B. Schneider, M. Wagner, M. Verhoeven, M. Bohm, "Vertically integrated sensors for advanced imaging applications", IEEE Journal of Solid-State Circuits, vol. 35, n° 7, July 2000, pp. 939-945.*

Bernard-97

→ *T.M. Bernard, "Rétines artificielles : quelle intelligence au sein du pixel ?", Hermes Paris, Calculateurs parallèles, special issue on {FPGAs ans Smart Sensors}, Mars 1997, pp. 77-108.*

Bertero et al.-88

→ *M. Bertero, T.A Poggio, V. Torre, "Ill-posed problems in early vision," in Proceedings of the IEEE, Vol. 8, 1988, pp. 869-889.*

Boahen-96

→ *K.A. Boahen, ,"Retinomorphic vision systems", in Proc. Of 5th Int. Conf. Microelectronics for Neural Networks and Fuzzy Systems, MicroNeuro'96, 1996, pp. 2-14.*

Borst & Egelhaaf-89

→ *A. Borst and M. Egelhaaf, "Principles of visual motion detection", Trends in Neuroscience, vol. 12, 1989, pp. 297-306.*

Bovik-00

→ *A.C. Bovik, "Handbook of image and video processing", Academic Press, ISBN 0-1211-9790-5, June 2000.*

BroadWare-05

→ *"Using Mpeg-4 technology in networked video surveillance systems", White paper, BroadWare, 2005.*

Burkey et al.-84

➙ *B.C. Burkey, W.C. Chang, J. Littlehale, T.H. Lee, T.J. Tredwell, J.P. Lavine, E.A. Trabka, “The pinned photodiode for an interline-transfer CCD image sensor”, Int. Electron Devices Meeting, Vol. 30, 1984, pp. 28-31.*

Burt-81

➙ *D. C. Burr, “Temporal summation of moving images by the human visual system”, Proc. Roy. Soc. B211, 1981, pp. 321-339.*

Burt-88

➙ *P.J. Burt, "Smart sensing within a pyramid vision machine", Proceeding of the IEEE, Special issu on computer vision, vol. 76, n° 8, August 1988, pp. 1006-1015.*

Burt & Adelson-83

➙ *P.J Burt and E.H. Adelson, "The Laplacian pyramid as a compact image code", IEEE trans. On Communications, vol. 31, n° 4, April 1983, pp. 532--540.*

Canon-05

➙ *http://web.canon.jp/Imaging/cmos/index-e.html.*

Cathebras et al.-02

➙ *G. Cathebras, D. Navarro, O. Aubreton, B. Bellach, P. Gorria, B. Lamalle, L.F.C. Lew, Yan Voon, "A continuous time pattern recognition retina", ; Proceedings of the European Solid-State Circuits Conference, ESSCIRC 2002, Sept. 2002, pp.719 - 722.*

Censi et al.-99

➙ *A. Censi, A. Fusiello, V. Roberto, "Image stabilization by features tracking", in Proc. of the 13th International Conference on Image Analysis and Processing, 27-29 Sept. 1999, pp. 665-667.*

Chung et al.-04

➙ *Y. Y. Chung, Y. Sun, P. Wang, and X. Chen, “A new e-shopping system using high performance custom computers,” in Proceedings of the 28th Computer Software and Applications Conference COMPSAC, vol. 2, Sept. 2004, pp. 54–55.*

CIE-04

➙ *Commission Internationale de l’Eclairage, "Colorimetry", 3rd Edition, Publication CIE 15:2004, ISBN 3-901-906-33-9, 2004.*

Clapp and Etienne-Cumming-02

➙ *M.A. Clapp and R. Etienne-Cummings, "A dual-pixel type array for imaging and motion centroid localization", IEEE Sensors Journal, vol. 2, n° 6, Dec. 2002, pp. 529-548.*

Cohen & Herlin-96

➙ *I. Cohen and I. Herlin, ” Non Uniform Multiresolution Method for Optical Flow and Phase Portrait Models: Environmental Applications", Institut National de Recherche en Informatique et en Automatique, Rapport de Recherche, No. 2819, March 1996, ISSN 0249-6399.*

Comby-01

➙ *F. Comby, ”Estimation du mouvement apparent majoritaire dans une séquence d’images vidéo par accumulation de votes bimodaux sur un histogramme approché", Thèse de doctorat, LIRMM, Soutenue le 19 décembre 2001.*

Corpetti et al.-00

➙ *T. Corpetti, E. Mémin, P. Pérez, "Dense Estimation of Fluid Flows", IEEE Transactions on Pattern Analysis and Machine Intelligence, Mars 2002, vol. 24, n° 3. pp 365-380.*

Degerli-00

➙ *Y. Degerli, "Etude, modélisation des bruits et conception des circuits de lecture dans les capteurs d’images à pixels actifs CMOS", Thèse de doctorat de l’Ecole Nationale Supérieure de l’Aéronautique et de l’Espace, 2000.*

Dekeyser et al.-00

➙ *F. Dekeyser, P. Bouthemy, P. Perez, E. Payot, "Super-resolution from noisy image sequences exploiting a 2D parametric motion model ", in Proceedings of 15th International Conference on Pattern Recognition, vol. 3, 3-7 Sept. 2000, pp. 350-353.*

Delbruck

➙ *http://www.ini.unizh.ch/~tobi/*

Delbruck-00

➙ *T. Delbruck, "Silicon retina for autofocus", IEEE International Symposium on Circuits and Systems, ISCAS 2000, Geneva, May 2000, vol. 4, pp.:393-396.*

Delbruck & Mead-89

➙ *T. Delbruck and C. Mead, "An electronic photoreceptor sensitive to small changes in intensity", Advances in Neural Information Processing Systems 1, D.S. Touretzky, Ed., Morgan Kaufman, San Mateo CA, 1989, pp. 720-726.*

Delbruck & Mead-91

➙ *T. Delbrück and C.A. Mead, "Time-derivative adaptive silicon photoreceptor array", In T.S. Jay Jayadev, SPIE, Infrared Sensors: Detectors, Electronics, and Signal Processing, vol. 1541, 1991, pp. 92-99.*

Delbruck & Mead-96

➙ *T. Delbruck and C. Mead, "Analog VLSI phototransduction", California Institute of Technology Computation and Neural Systems Program, CNS Memo No. 30, April 2, 1996.*

Delbruck & Liu-04

➙ *T. Delbruck and S-C. Liu, ” A silicon early visual system as a model animal ", in Vision Research, Vol. 44, n° 17, 2004, pp. 2083-2089.*

Delbruck & Oberhof-04

➙ *T. Delbruck, D. Oberhof, "Self biased low power adaptive photoreceptor", IEEE International Symposium on Circuits and Systems, ISCAS 2004, Vancouver, May 2004, Vol. 4, pp. 844-347.*

Deutschmann & Koch-98a

➙ *R.A. Deutschmann and C. Koch, “An analog VLSI velocity sensor using the gradient method”, IEEE Int. Symp. on Circuits and Systems, Vol. 6, 31st May- 3rd June, 1998, pp. 649-652.*

Deutschmann & Koch-98b

➙ *R. A. Deutschmann and C. Koch, “Compact real-time 2-D gradient based analog VLSI motion sensor”, Int. Conf. Advanced Focal Plane Arrays and Electronic Cameras, Zurich/Switzerland, 1998.*

Dierickx et al.-97

➙ *B. Dierickx, G. Meynants, and D. Scheffer, “Near 100% fill-factor standard CMOS active pixel”, IEEE Workshop on Charge-Coupled Devices and Advanced Image Sensors, 1997.*

DigitalPhotography-06

➙ *Digital Photography Review Newsletter, .*

Dong et al.-05

➙ *L. Dong, R. Yue, L. Liu, "Fabrication and Characterization of Integrated Uncooled Infrared Sensor Arrays Using a-Si Thin-Film Transistors as Active Elements", Journal of Microelectromechanical Systems, vol. 14, n° 5, October 2005, pp. 1167-1177.*

Dubois-85

➙ *R. Dubois, "The sampling and reconstruction of time-varying imagery with application in video systems", IEEE Proc., 1985, vol. 73, n° 4, pp. 502-522.*

Dudek & Hicks-05

➙ *P. Dudek and P.J. Hicks, “A general-purpose processor-per-pixel analog SIMD vision chip”, IEEE Trans. Circuits and Systems I, Vol. 52, n° 1, January 2005, pp. 13-20.*

Dufaux & Moscheni-95

➙ *F. Dufaux and F. Moscheni, "Motion estimation techniques for digital TV: a review and a new contribution", Proc. IEEE, vol. 83, 1995, pp. 858-876.*

Dufaux & Moscheni-96

➙ *F. Dufaux and F. Moscheni, "Segmentation-based estimation for second generation video coding techniques", in Video Coding: The Second Generation Approach, L. Torres and M. Kunt, Eds. Boston, MA: Kluwer, 1996, pp. 219--264.*

Dupret et al.-02

➙ *A. Dupret, J. O. Klein, A. Nshare, “A DSP-like analog processing unit for smart image sensors”, Int. J. Circuit Theory Applicat., vol. 30, 2002, pp. 595–609.*

Dutta-03

➙ *A. Dutta, "An image stabiliser for a microcamera module of a handheld device and method for stabilizing a microcamera module of a hand-held device", Nokia Corp., Pub. date: April 23th 2003, Patent N°: EP1304872.*

Elad & Feuer-97

➙ *M. Elad and A. Feuer, “Restoration of Single Super-Resolution Image From Several Blurred, Noisy and Down-Sampled Measured Images”, IEEE Trans. on Image Processing, vol. 6, n°. 12, December 1997, pp. 1646-58.*

Elect. Int.-05

➙ *Electronique International Hebdo, 10 Février 2005.*

Elouardi-05

➙ *A. Elouardi, “Évaluation des rétines électroniques pour une définition architecturale d’un système monopuce (SoC) dédié à la vision embarquée », Thèse soutenue le 25/05/2005 à Insitut d'Electronique Fondamentale, Université Paris Sud, 2005.*

Engelhardt et al.-92

➙, *K. Engelhardt, J. Kramer, D. Leipold, J.M. Raynor, P. Seitz, E. Tan, "Smarter Sensor Pixels Through Geometry Control", Broadband Analog and Digital Optoelectronics, Optical Multiple Access Networks, Integrated Optoelectronics, Smart Pixels, LEOS 1992 Summer Topical Meeting Digest on 29 July-12 Aug. 1992, pp. C7-C8.*

Engelsberg & Schmidt-99

➙ *A. Engelsberg and G. Schmidt, "A comparative review of digital image stabilising algorithms for mobile video communications", IEEE Transactions on Consumer Electronics, Aug. 1999, vol. 45, n° 3, pp. 591-597.*

Etienne-Cummings et al.-93

➙ *R. Etienne-Cummings, S. Fernando, N. Takahashi, V. Shtonov, J. Van der Spiegel, P. Muller, “A new temporal domain optical flow measurement technique for focal plane VLSI implementation”, Computer Architectures for Machine Perception, 15-17 Dec. 1993, pp. 241-250.*

Etienne-Cummings et al.-00

➙ *R. Etienne-Cummings, J. Van der Spiegel, P. Mueller and M.Z. Zhang, “A Foveated Silicon Retina for Two-Dimensional Tracking”, IEEE Trans. Circuits and Systems II, Vol. 47, June 2000, pp. 504-517.*

Etienne-Cummings et al.-01

➙ *R. Etienne-Cummings, "Neuromorphic Visual Motion Detection in VLSI", Int. J. Computer Vision, Vol. 44, No. 3, Sept. 2001, pp. 175-198.*

Fermuller & Aloimonos-98

➙ *C. Fermuller and Y. Aloimonos, ”Geometry of eye design : biology and technology", Maryland Univ. Technical Report, CS-TR-3963, December 1998.*

Findlater et al.-03a

➙ *K.M. Findlater, D. Renshaw, J.E.D. Hurwitz, R.K. Henderson, M.D. Purcell, S.G. Smith, T.E.R. Bailey, "A CMOS image sensor with a double-junction active pixel", IEEE Transaction on Electronic Devices, vol. 50, n° 1, January 2003, pp. 32-42.*

Findlater et al.-03b

➙ *K.M. Findlater, R. Henderson, D. Baxter, J.E.D. Hurwitz, L. Grant, Y. Cazaux, F. Roy, D. Herault, Y. Marcellier, "SXGA pinned photodiode CMOS image sensor in 0.35μm technology", IEEE International Solid-State Circuits Conference, 2003, vol. 1.*

Fisher et al.-92

➙ *H. Fischer, J. Schulte, J. Giehl, M. Böhm, J. P. M. Schmitt, "Thin film on ASIC - a novel concept for intelligent image sensors", Mat. Res. Soc. Symp. Proc., Vol. 285, 1992, pp. 1139-1145.*

Fleet & Jepson-89

➙ *D.J. Fleet and A.D. Jepson "Computation of formal velocity from local phase information", Computer Vision and Pattern Recognition, Proceeding IEEE, June 1989, pp. 379 – 386.*

Fossum-97

➙ *E.R. Fossum, "CMOS image sensors : electronic-camera-on-a-chip", IEEE Transactions on Electron Devices, vo. 44, n° 10, Oct. 1997, pp. 1689-1698.*

Franceschini-99

➙ *N. Franceschini, "De la mouche au robot : Reconstruire pour mieux comprendre", In V. Bloch, editor, Cerveaux et Machines, Hermes, Paris,1999, pp. 247-270.*

Froba & Ernst-04

➙ *B. Froba and A. Ernst, "Face detection with the modified census transform," in Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, May 2004, pp. 91-96.*

Fujisaki et al.-01

➙ *T. Fujisaki, I. Nakazawa, Y. Shiomi, "Image stabilizing device operable responsively to a state of optical apparatus using the same", Canon KK, Pub. date: Feb. 20th 2001, Patent N°: US6191813.*

Funatsu et al.-97

➙ *E. Funatsu, Y. Nitta, Y. Miyake, T. Toyoda, J. Ohta, and K. Kyuma, "An artificial retina chip with current-mode focal plane image processing functions", IEEE Trans. On Electron Devices, vol. 44, n° 10, Oct. 1997, pp. 1777-1782.*

Furukawa & Tajima-76

➙ *H. Furukawa and A. Tajima, "Image stabilizing optical system", Canon KK, Pub. date: April 27th 1976, Patent n° US3953106.*

Future Horizons-05

➙ *Future Horizons, "Semiconductor application markets report", Ed. 2005.*

Giese & Poggio-03

➙ *M.A. Giese and T. Poggio, "Neural mechanisms for the recognition of biological movements", Nature Neuroscience Review, vol. 4, March 2003, pp. 179-192.*

Golston-04

➙ *J. Golston, "Comparing media codecs for video content", Embedded Systems Conference, San Francisco 2004, Paper #250, www.ti.com.*

Green et al.-04

➙ *W.E. Green, P.Y. Oh, G. Barrows, "Flying insect inspired vision for autonomous aerial robot maneuvers in near-earth environments", IEEE Int. Conf. on Robotic and Automation, New Orleans, LA, April 2004, pp. 2347-2352.*

Grossberg et al.-00

➙ *S. Grossberg, E. Mingolla, L. Wiswanathan, "Neural dynamics of motion integration and segmentation within and across apertures", Bolston Univ. Technical Report, CAS-CNS-2000-004, January 2000.*

Gruev & Etienne-Cummings-04

➙ *V. Gruev and R. Etienne-Cummings, "A pipelined temporal difference imager", IEEE Journal of Solid State Circuits, Vol. 39, No. 3, March 2004, pp. 538-543.*

Hadamard-1902

➙ *J . Hadamard, "Sur les problèmes aux dérivées partielles et leur signification physique", Princeton University Bulletin, vol. 13, 1902.*

Hager & Belhumeur-98

➙ *G. Hager and P. Belhumeur, "Efficient region tracking with parametric models of geometry and motion", IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 20, No. 10, Oct. 1998, pp. 1025-1039.*

Harrison & Koch-00

➙ *R.R. Harrison and C. Koch, "A robust analog VLSI Reichardt motion sensor", Computation and Neural Systems Program, ALOG868-99, California Institute of Technology, Pasadena CA 91125 USA, 8th February, 2000.*

Heeger-88

➙ *D.J. Heeger, "Optical flow using spatiotemporal filters", International Journal of Computer Vision, vol. 1, n° 4, 1988, pp.279-302.*

Higgins et al.-99

➙ *C.M. Higgins, R. Deutschmann, C. Koch, "Pulse-based 2-D motion sensors", IEEE Trans. on Circuits and Systems II., Vol. 46, No. 6, June 1999, pp. 677-687.*

Higgins et al.-05

➙ *C.M. Higgins, V. Pant, R. Deutschmann, "Analog VLSI implementation of spatio-temporal frequency tuned visual motion algorithms", IEEE Trans. on Circuits and Systems I., Vol. 52, No. 3, March 2005, pp. 489-502.*

Hildreth-84

➙ *E. Hildreth, "The measurement of visual motion",MIT Press, April 1984, ISBN: 0262081431.*

Horaud & Monga-95

➙ *R. Horaud and O. Monga, "Vision par ordinateur : outils fondamentaux", Hermes Ed., Second Edition, 1995.*

Horiuchi et al.-91

➙ *Horiuchi, T., Lazzaro, J., Moore, A. and Koch, C. "A delay-line based motion detection chip", Advances in Neural Information Processing Systems 3 Touretzky, D.S. and Lippman, R., eds., . Morgan Kaufmann, SanMateo, 1991, pp. 406-412.*

Horn & Schunck-80

➙ *B.K.P. Horn and B.G. Schunck, "Determining optical flow", Massachusetts Institute of Technology, Artificial Intelligence Laboratory, A.I. Memo No. 572, 1980.*

IC Insights-05

➙ *IC Insights' Emerging IC Markets Report 2005 details IC usage by electronic system type, Scottsdale, Arizona, January 19, 2005.*

Ishikawa et al.-99

➙ *M. Ishikawa, K. Ogawa, T. Komuro, and I. Ishii, "A CMOS vision chip with SIMD processing element array for 1 ms image processing", in Proc. Int. Solid State Circuits Conf., 1999, Paper No. TP 12.2.*

Jain-89

➙ *A.K. Jain, "Fundamentals of digital image processing", Prentice Hall, 1989.*

Johns & Martin-97

➙ *D. Johns and K. Martin, "Analog integrated circuit design", John Wiley & Sons, ISBN 0-471-14448-7, 1997.*

Jolion-01

➙ *J-M. Jolion, "Les systèmes de vision", Hermes Science Europe, ISBN 2-7462-0185-2, 2001.*

Kage et al.-98

➙ *H. Kage, K. Tanaka, K. Kyuma, "3-D human motion sensing by artificial retina chips", in Proceedings of the Third IEEE Int. Conf. on Automatic Face and Gesture Recognition, April 1998, pp. 522-527.*

Kay-93

➙ *S. Kay, "Fundamentals of Statistical Signal Processing: Estimation Theory", Prentice Hall, 1993.*

Karim et al.-03

➙ *K.S. Karim, A. Nathan, J.A. Rowlands, S.O. Kasap, " X-ray detector with on-pixel amplification for large area diagnostic medical imaging", IEE Proceedings on Circuits, Devices and Systems, 5 Aug. 2003, vol. 150, n° 4, pp. 267-273.*

Kasano et al.-05

➙ *M. Kasano, Y. Inaba, M. Mori, S. Kasuga, T. Murata, T. Yamaguchi, "A 2.0μm pixel pitch MOS image sensor with an amorphous Si film color filter", IEEE International Solid-State Circuits Conference, February 2005, pp. 348-350.*

Koch-91

➙ *C. Koch, "Implementing early visual algorithms in analog hardware: an overview", Visual Information Processing: from Neurons to chips, SPIE, Vol. 1473, 1991, pp. 2-16.*

Koch & Mathur-96

➙ *C. Koch and B. Mathur, "Neuromorphic vision chips", in IEEE Spectrum, Vol. 33, n° 5, May 1996, pp. 38-46.*

Koga et al.-81

➙ *T. Koga, K. Iinuma, A. Hirano, Y. Iijima, T. Ishiguro, "Motion-compensated interframe coding for video conferencing", IEEE Proceedings NTC'81, pp. G.5.3.1-G.5.3.4.*

Konishi-86

➙ *M. Konishi, "Centrally synthesized maps of sensory space", Trends in Neuroscience, No. 4, 1986, pp. 163-168.*

Kyuma et al.-97

➙ *K. Kyuma, Y. Miyake, H. Kage, "Artificial retina chips", Int. Conf. On Neural Networks, vol. 4, June 1997, pp. 2304-2308.*

Kramer-96

➙ *J. Kramer, "Compact integrated velocity sensor with three-pixel interaction", IEEE Trans. Patt. Anal. Mach. Intell., Vol. 18, Apr. 1996, pp. 455–460.*

Kramer-02

➙ *J. Kramer, "An Integrated Optical Transient Sensor", IEEE Trans. on Circuits and System II, Analog and Digital Signal Processing, Vol. 49, No. 9, Sept. 2002, pp. 612-628.*

Kramer et al.-94

➙ *J. Kramer, P. Seitz, H. Baltes, "Planar distance and velocity sensor", IEEE Journal of Quantum Electronics, vol. 30, n° 11, Nov. 1994, pp. 2726-2730.*

Kramer et al.-96

➙ *J. Kramer, R. Sarpeshkar, C. Koch, " Analog VLSI motion discontinuity detectors for image segmentation ", IEEE Int. Symp. Circuits Syst. II, , 1996, pp. 620-623.*

Kramer & Indiveri-96

➙ *J. Kramer, G. Indiveri and C. Koch, "Analog VLSI motion projects at Caltech", in T. M. Bernard (ed.), Advanced Focal Plane Arrays and Electronic Cameras (AFPAEC '96), Berlin, Germany, Oct. 9-10, 1996, Proc. SPIE 2950, SPIE, Bellingham, WA, USA, 1996, pp. 50-63.*

Kuhn & Stechele-98

➙ *P. M. Kuhn and W. Stechele, "Complexity analysis of the emerging MPEG-4 standard as a basis for VLSI implementation", in Proceedings of the Int. Conf on Visual Communications and Image Processing, San Jose, Jan. 1998, vol. SPIE 3309, pp. 498-509.*

Kyuma-99

➙ *K. Kyuma, "Concept, development, mass production, and applications of artificial retina chips", Int. Conf. on Intelligent Processing and Manufacturing of materials, vol. 2, July 1999, pp. 1297-1303.*

Lin et al.-02

➙ *C-S.S. Lin, B.P. Mathur, M.C.F. Chang, "Analytical charge collection and MTF model for photodiode-based CMOS imagers", IEEE Trans. on Electronic Devices, vol. 49, n° 5, May 2002, pp. 754-761.*

Liu -00

➙ *S-C. Liu, "A Neuromorphic aVLSI Model of Global Motion Processing in the Fly ", IEEE Trans. On Circuits and Systems II, Analog and Digital Signal Processing, Vol. 47, No. 12, Dec. 2000..*

Loose et al.-01

➙ *M. Loose, Karlheintz Meier, Johannes Schemmel, "A self-calibrating single-chip CMOS camera with logarithmic response", IEEE Journal of Solid-State Circuits, vol. 36, n° 4, April 2001, pp. 586-596.*

Lu et al.-01

➙ *G.N. Lu, G. Sou, F. Devigny, G. Guillaud, "Design and testing of a CMOS BDJ detector for integrated micro-analysis systems", in Microelectronics Journal, Elsevier, n° 32, 2001, pp. 227-234.*

Lu et al.-03

➙ *L. Lu, X-T. Dai, G. Hager, "Real-time Video Mosaicing with Adaptive ParametricWarping, Computational Interaction and Robotics Lab Computer Science Department the Johns Hopkins University Baltimore, MD 21218, USA, Technical Report 2003-01-CIRL-CS-JHU, 2003.*

Lule et al.-99

➙ *T. Lule, B. Schneider, M. Bohm, "Design and fabrication of a high-dynamic-range image sensor in TFA technology", IEEE Journal of Solid-State Circuits, vol. 34, n° 5, May 1999, pp. 704-711.*

Lule et al.-00a

➙ *T. Lule, S. Benthien, H. Keller, F. Mutze, P. Rieve, K. Seibel, M. Sommer, M. Bohm, "Sensitivity of CMOS based imagers and scaling perspectives", IEEE Trans. on Electronic Devices, vol. 47, n° 11, November 2000, pp. 2110-2122.*

Lule et al.-00b

➙ *T. Lule, M. Wagner, M. Verhoeven, H. Keller, M. Bohm "100 000-pixel, 120dB imager in TFA technology", IEEE Journal on Solid State Circuits, vol. 35, n° 5, May 2000, pp. 732-739.*

Lyons-01

➙ *R. Lyons, "Understanding digital signal processing", Prentice Hall, ISBN 0-201-63467-8, 2001.*

Magnan-03

➙ *P. Magnan, "Detection of visible photons in CCD and CMOS: A comparative view", Journal of Nuclear Instruments and Methods in Physics Research A, May 2003, vol. 504, pp. 199-212.*

Magnan et al.-04

➙ *P. Magnan, M. Estribeau, C. Marques, S. Maëstre, J-M. Baqué, "Amélioration et modélisation des performances des capteurs d'images CMOS", www.chear.defense.gouv.fr, RSTD, n° 63, Mars-Avril 2003, pp. 71-86.*

Mahowald & Mead-89

➙ *M.A. Mahowald and C.A. Mead, "Silicon retina", in Mead C.A. Analog VLSI and Neural Systems, Addison-Wesley, 1989, pp. 267-278.*

Mathieu-96

➙ *H. Mathieu, “Physique des semiconducteurs et des composants électroniques", Masson, 3e édition, ISBN 2-225-85124-7, 1996.*

Matou-03

➙ *K. Matou, "Capteurs d’images logarithmiques CMOS avec compensation « on-chip » du bruit spatial fixe", Thèse de doctorat, Université Parix VI Orsay, 2003..*

Mead-89

➙ *C. Mead, “Analog VLSI and Neural Systems”, Addison Wesley, New York, 1989.*

Mendis et al.-93

➙ *S.K. Mendis, S.E. Kemeny, E.R. Fossum, “A 128×128 CMOS active pixel image sensor for highly integrated imaging systems”, Int. Electron Devices Meeting, Technical Digest., 5-8 Dec. 1993, pp. 583-586.*

Miller & Barrows-99

➙ *K.T. Miller, G.L. Barrows, “Feature tracking linear optic flow sensor chip”, Int. Symp. on Circuits and Systems, Vol. 5, 30th May-2nd June 1999, pp. 116-119.*

Mitiche & Bouthemy-96

➙ *A. Mitiche and P. Bouthemy, ”Computation and analysis of image motion : a synopsis of current problems and methods", International Journal of Computer Vision, vol. 19 , n° 1, July 1996, pp. 29-55.*

Moini-99

➙ *A. Mioni, "Vision chips", Kluwer Academic Publishers, ISBN 0-7923-9664-7, 1999.*

Moini et al.-93

➙ *A. Moini, A. Bouzerdoum, A. Yakovleff, D. Abbott, O. Kim, K., Eshraghian, and R. E. Bogner, “An analog implementation of early visual processing in insects,” in Proc. 1993 Int. Symp. VLSI Technology, Systems and Applications, Taipei, Taiwan, May 1993, pp. 283–287.*

Moini et al.-97

➙ *A. Moini, A. Bouzerdoum, K. Eshraghian, A. Yakovleff, X. T. Nguyen, A. Blanksby, R. Beare, D. Abbott, and R. E. Bogner, “An Insect Vision-Based Motion Detection Chip”, IEEE Journal of Solid-State Circuits, Vol. 32, No. 2, February 1997, pp. 279-284.*

Montesinos

➙ *P. Montesinos, “Détection de contours”, LGI²P, EMA-ERIEE, Parc Scientifique G. Besse, Nîmes.*

Mori et al.-04

➙ *M. Mori, M. Katsuno, S. Kasuga, T. Murata, T. Yamagushi, "1/4-inch 2-Mpixel MOS image sensor with 1.75 transistors/pixel", IEEE Journal of Solid-State Circuits, vol. 39, n° 12, December 2004, pp. 2426-2430.*

Morimoto & Chellappa-96

➙ *C. Morimoto and R. Chellappa, "Fast electronic digital image stabilization", in Proc. of the 13th International Conference on Pattern Recognition, 25-29 Aug. 1996, vol. 3, pp. 284-288.*

Navarro-03

➙ *D. Navarro, “Architecture et conception de rétines silicium CMOS : application à la mesure du flot optique’’, Thèse de doctorat, Université Montpellier II, 2003.*

Ni et al.-93

➙ *Yang Ni; Yi-min Zhu; B. Arion, F. Devos, “Yet Another Analog 2D Gaussian Convolver”, Int. Symp. on Circuits and Systems, 3-6 May 1993, pp. 192-195.*

Ni et al.-96

➙ *Y. Ni, F. Devos, B. Arion, "Analog retina based real-time vision system", Int. Conf. on Semiconductor, vol. 1, 9-12 Oct. 1998, pp. 275-284.*

Odobez & Bouthemy-95

➙ *J.M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models", Journal of Visual Communication and Image Representation, December 1995, 6(4), pp.348-365.*

Oike et al.-04

➙ *Y. Oike, M. Ikeda, K. Asada, "Design and implementation of real-time 3-D image sensor with 640 /spl times/ 480 pixel resolution", IEEE Journal of Solid-State Circuits, vol. 39, n° 4, April 2004, pp. 622-628.*

Olveczky et al.-03

➙ *B.P. Olveczky, S.A. Baccus, M. Meister, ”Segregation of object and background motion in the retina", in Nature, May 2003, pp. 401-408.*

Paillet et al.-98

➙ *F. Paillet, D. Mercier, T. M. Bernard, “Making the most of 15kλ² silicon area for a digital retina”, in Proc. SPIE, vol. 3410, 1998, pp.158–167.*

Pain et al.-01

➙ *B. Pain, S. Seshadri, M. Ortiz, C. Wrigley, G. Yang, “CMOS imager with charge-leakage compensated frame difference and sum output”, in Proc. Int. Symp. On Circuits and Systems, vol. 5, May 2001, pp. 223-226.*

Pardo et al.-97

➙ *F. Pardo, B. Dierickx, D. Scheffer, "CMOS foveated image sensor: signal scaling and small geometry effects", IEEE Transactions on Electron Devices, vol. 44, n°. 10, Oct. 1997, pp. 1731-1737.*

Weckler-67

➙ *G. P.Weckler, “Operation of p-n junction photodetectors in a photon flux integration mode,” IEEE Journal of Solid-State Circuits, vol. SC-2, 1967, pp. 65–73.*

Park et al.-03

➙ *J-H. Park, J-H. Kim, J-K. Shin, L. Lee, P. Choi, “Edge and motion detection using a bio-inspired CMOS vision chip robust to device mismatches”, IEEE Int. Conf.on Neural Networks and Signal Processing, Nanjing, China, December 14-17, 2003, pp. 341-344.*

Perrinet-02

➙ *L. Perrinet, ”Comment déchiffrer le code impulsionnel de la vision ? Etude du flux parallèle, asynchrone et épars dans le traitement visuel ultra-rapide", Thèse de doctorat, Université Paul Sabatier, 2002.*

Rahimi et al.-05

➙ *M. Rahimi ,R. Baer ,O.I. Iroezi , J.C. Garcia ,J. Warrior ,D. Estrin ,M. Srivastava, "Cyclops: In Situ Image Sensing and Interpretation in Wireless Sensor Networks", To appear in the Third ACM Conference on Embedded Networked Sensor Systems (SenSys), November 2-4, 2005.*

Reichardt-61

➙ *W. Reichardt, ”Autocorrelation, a principle for the evaluation of sensory information by the central nervous system", in Principles of sensory communication, W. A. Rosenblith, ed., Wiley, New York, 1961, pp. 303-317.*

Rhodes et al.-04

➙ *H. Rhodes, G. Agranov, C. Hong, U. Boettiger, R. Mauritzson, J. Ladd, I. Karasev, J. McKee, E. Jenkins, W. Quinlin, I. Patrick, J. Li, X. Fan, R. Panicacci, S. Smith, C. Mouli, J. Bruce, "CMOS imager technology shrinks and image performance", IEEE Workshop on Microelectronics and Electron Devices, 2004, pp. 7-18.*

Rodriguez-Vazquez et al.-04

➙ *A. Rodríguez-Vázquez, G. Liñán-Cembrano, L. Carranza, E. Roca-Moreno, R. Carmona-Galán, F. Jiménez-Garrido, R. Domínguez-Castro, and S. Espejo Meana, “ACE-16k: the third generation of mixed-signal SIMD-CNN ACE chips towards VSoCs”, IEEE Trans. on Circuits and Systems I, Regular Papers, Vol. 51, No. 5, May 2004, pp. 851-863.*

Ruedi-96

➙ *P.F. Ruedi, ,"Motion detection silicon retina based on event correlations", in Proc. Of 5th Int. Conf. Microelectronics for Neural Networks and Fuzzy Systems, MicroNeuro’96, 1996, pp. 23-29.*

Ruedi et al.-03

➙ *P.F. Ruedi, P. Heim, F. Kaess, E. Grenet, F. Heitger, P.Y. Burgi, S. Gyger, P. Nussbaum, "A 128×128 pixel 120-dB dynamic range vision-sensor chip for image contrast and orientation extraction", IEEE Journal of Solid-State Circuits, vol. 38, n° 12, December 2003, pp. 2325-2333.*

Ryer-97

➙ *Alexander Ryer, "The light measurement handbook", 2nd Printing, International Light, ISBN 0-9658356-9-3, 1997.*

Sandini et al.-00

➙ *G. Sandini, P. Questa, D. Scheffer, B. Dierickx, A. Mannucci, "A retina-like CMOS sensor and its applications", Proceedings of the IEEE Sensor Array and Multichannel Signal Processing Workshop, 2000, 16-17 March 2000, pp. 514-519.*

Schreer et al.-01

➙ *O. Schreer, N. Brandenburg, C. Plakas, E. Trucco, and P. Kauff, "A combination of census transform and a hybrid block-pixel-recursive disparity analysis approach for real-time videoconferencing applications," in Proceedings of ICAV3D, Mykonos, Greece, May 2001, pp. 29-32.*

Shcherback & Yadid-Pecht-04

➙ *I. Shcherback and O. Yadid-Pecht, " Prediction of CMOS APS design enabling maximum photoresponse for scalable CMOS technologies ", IEEE Trans. On Electron Devices, vol. 51, n° 2, Feb. 2004, pp. 285-288.*

Simpson et al.-99

➙ *M.L. Simpson, N.N. Ericson, G.E. Jellison, W.B. Dress, A.L. Wintenberg, M. Bodreck, "Application specific spectral response with CMOS compatible photodiodes", IEEE Trans. on Electronic Devices, vol. 46, n° 5, May 1999, pp. 905-913.*

Smolic et al.-00

➙ *A. Smolic, M. Hoeynck, J-R. Ohm, "Low-complexity global motion estimation from P-frame motion vectors for MPEG-7 applications", in Proc. of the 13th International Conference on Image Processing, Sept. 2000, vol. 2, pp. 271-274.*

SNOF

➙ *Syndicat National des Ophtalmologistes de France, "La vision des animaux", http://www.snof.org/vue/animaux.html.*

Song & Chun-04

➙ *B.C. Song and K-W Chun, " Multi-Resolution Block Matching Algorithm and Its VLSI Architecture for Fast Motion Estimation in an MPEG-2 Video Encoder", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 14, No°. 9, September 2004, pp. 1119-1137.*

Spindler & Bouthemy-98

➙ *F. Spindler, P. Bouthemy, "Real-time estimation of dominant motion in underwater video images for dynamic positionning", IEEE Int. Conf. on Robotics and Automation, ICRA'98, May 1998, Leuven, Belgium, vol. 2, pp. 1063-1068.*

Spinei et al.-98

➙ *A. Spinei, D. Pellerin, J. Herault, "Spatiotemporal Energy-Based Method for Velocity Estimation", Signal Processing: an International Journal, vol. 65, 1998, pp. 347-362.*

SST-04

➙ *J. Borland, N. Tokoro, Solid State Technology, November 2004.*

Stiller & Konrad-99

➙ *C. Stiller and J. Konrad, "Estimating motion in image sequences: A tutorial on modeling and computation of 2D motion", IEEE Signal Process. Mag., July 1999, vol. 16, pp. 70-91.*

Stocker-04

➙ *A. Stocker, "Optical flow estimation as distributed optimization problem - an aVLSI Implementation", Tech. Report TR2004-850, Computer Science Dept., New York University, January 2004.*

Stocker-06

➙ *A. Stocker, "Analog integrated 2-D optical flow sensor", Analog Integrated Circuits and Signal Processing, Springer Science, Vol. 46, 2006, pp. 121-138.*

Stoppa et al.-02

➙ *D. Stoppa, A. Simoni, L. Gonzo, M. Gottardi, G. F. D. Betta, "Novel CMOS image sensor with a 132-dB dynamic range," IEEE Journal of Solid-State Circuits, vol. 37, n° 12, Dec. 2002, pp. 1846-1852.*

Sutherland et al.-02

➙ *Sutherland, A.J.; Hamilton, A.; Renshaw, D.A.; Glover, M.A., "Analogue VLSI for temporal frequency analysis of visual data", IEEE International Symposium on Circuits and Systems, ISCAS 2002, May 2002, vol. 3, pp. 743-746.*

Sze-81

➙ *S.M. Sze, "Physics of semiconductor devices", 2nd Edition, Wiley, ISBN 0-471-05661-8, 1981.*

Theil-03

➙ *A.J. Theil, " Advances in elevated diode technologies for integrated circuits: progress towards monolithic instruments", IEE Proceedings.of Circuits, Devices and Systems, 5 Aug. 2003, vol. 150, n° 4, pp. 235-249.*

Tan et al.-00

➙ *Y-P. Tan, D.D. Saur, S.R. Kulkami, P.J. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation", IEEE Trans. on Circuits and Systems for Video Technology, Feb. 2000, vol. 10, n° 1, pp. 133-146.*

Tanner & Mead-84a

➙ *J. Tanner and C. Mead, "Correlating optical motion detector", California Institute of Technology, Filed date: Jan. 20, 1984, Patent n° 4631400.*

Tanner & Mead-84b

➙ *J. Tanner and C. Mead, "A correlating optical motion detector", In Pentfield, P. (ed), Proc. Of Conf. on Advanced Research in VLSI, MIT, Cambridge, January 23-25, 1984, pp. 57.*

Tanner & Mead-86

➙ *J. Tanner and C. Mead, "An integrated analog optical motion sensor", in VLSI Signal Processing, II, S.Y. Kung, R. Owen, and G. Nash (Eds.), New York : IEEE Press, pp. 59-76, 1986.*

Theuwissen-02

➙ *A.J.P. Theuwissen, "Small is beautiful ! Yes, but also for pixels of digital still cameras ?", International Technical Conference on Digital Image Capture and Associated System,Reproduction and Image Quality Technologies, PICS 2002, Portland, Oregon, April 2002, pp. 156-157.*

Tian et al.-01

➙ *H. Tian, X. Liu, S. Lim, S. Kleinfelder, A. El Gamal, "Active pixel sensors fabricated in a standard 0.18um CMOS technology," Proceedings of SPIE, January 2001, Vol. 4306, pp. 441-449.*

Tirunelveli et al.-02

➙ *G. Tirunelveli, R. Gordon, S. Pistorius, "Comparison of square-pixel and hexagonal-pixel resolution in image processing", IEEE Canadian Conf. On Electrical and Computer Engineering, Vol. 2, 12-15 May 2002, pp. 867-872.*

Torralba & Herault-99

➙ *A. B. Torralba and J. Herault, "An efficient neuromorphic analog network for motion estimation", IEEE Trans. on circuits and systems-I: special issue on bio-inspired processors and CNNs for vision, Vol. 46, Feb. 1999, pp. 269-280.*

Tremblay et al.-93

➙ *M. Tremblay, M. d'Anjou, D. Poussart, "Hexagonal sensor with embedded analog image processing for pattern recognition", Proceedings of the IEEE Custom Integrated Circuits Conference, 9-12 May 1993, pp. 12.7.1-12.7.4.*

Tremblay et al.-95

➙ *M. Tremblay, M. d'Anjou, D. Poussart, " Low level segmentation using CMOS smart hexagonal image sensor", Proc. of Computer Architectures for Machine Perception, 18-20 Sept. 1995, pp. 21-28.*

Umminger & Sodini-92

➙ *C.N. Umminger and C.G. Sodini, "Switched capacitor networks for focal plane image processing systems", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 2, No. 4, Dec. 1992, pp. 392-400.*

Uomori et al.-90

➙ *K. Uomori, A. Morimura, H. Ishii, T. Sakaguchi, Y. Kitamura, "Automatic image stabilizing system by full-digital IEEE Transactions on signal processing", Aug. 1990, Consumer Electronics, vol. 36, n° 3, pp.510-519.*

Vaillant & Hirigoyen-04

➙ *Jerome Vaillant and Flavien Hirigoyen, "Optical simulation for CMOS imager microlens optimization", in proc. SPIE Optical Sensing Eds., Sept. 2004, vol. 5459, p. 200-210.*

Van der Spiegel et al.-89

➙ *J. Van der Spiegel, G. Kreider, C. Claeys, I. Debusschere, G. Sandini,P. Dario, F. Fantini, P. Belluti, G. Soncini, "A Foveated Retina-Like Sensor Using CCD Technology", eds. C. Mead and M. Ismail, Kluwer Academic Publ., chap. 8, 1989, pp. 189-210.*

Vella et al.-02

➙ *F. Vella, A. Castorina, M. Mancuso, G. Messina, "Digital image stabilization by adaptive block motion vectors filtering", IEEE Trans. On Consumer Electronics, Vol. 48, No.3, August 2002, pp. 796-801.*

Viarani et al.-04

➙ *L. Viarani, D. Stoppa, L. Gonzo, M. Gottardi, A. Simoni, "A CMOS smart pixel for active 3-D vision applications", IEEE Sensors Journal, vol. 4, n° 1, Feb. 2004, pp. 145-152.*

Video/Imaging DesignLine-06

➙ *J. Yoshida, "Camera cell phone designers challenged to improve image quality. The need to improve low-end sensors has become a rallying cry for designers", 27th March, 2006.*

Vittoz-94

➙ *E. Vittoz, "Analog VLSI signal processing : why, where and how ?", Analog Integrated Circuits and Signal Processing, vol. 6, n° 1, 1994, pp. 27-44.*

Vittoz & Arreguit-93

➙ *E. Vittoz and X. Arreguit, "Linear networks based on transistors", Electronics Letters, Vol. 29, No. 3, 4th February 1993, pp. 297-299.*

Weckler-67

➙ *G. P.Weckler, "Operation of p-n junction photodetectors in a photon flux integration mode," IEEE Journal of Solid-State Circuits, vol. SC-2, 1967, pp. 65–73.*

Weickert & Schnörr-01

➙ *J. Weickert, C. Schnörr, "Variational optic flow computation with a spatio-temporal smoothness constraint", Journal of Mathematical Imaging and Vision, Vol. 14, No. 3, 245-255, May 2001.*

Werblin-74

➙ *F. Werblin, "Control of retinal sensitivity II: lateral interactions at the outer plexiform layer", J. Physiology, 63, 1974, pp. 62-87.*

Wodnicki et al.-95

➙ *R. Wodnicki, G. W. Roberts, and M. D. Levine, "A foveated image sensor in standard CMOS technology," in Custom Integrated Circuits Conf., Santa Clara, CA, May 1995, pp. 357-360.*

Woodfill & Herzen-97

➙ *J. Woodfill, B. Von Herzen, "Real-time stereo vision on the PARTS reconfigurable computer", IEEE Symp. on FPGAs for Custom Computing Machines, 16-18 April, 1997, pp. 201-210.*

Wu et al.-04

➙ *C-Y. Wu, Y-C. Shih, J-F Lan, C-C. Hsieh, C-C. Huang, J-H Lu, "Design,optimization, and performance analysis of new photodiodes structures for CMOS active-pixel-sensor (APS) imager applications", IEEE Sensors Journal, vol. 4, n° 1, February 2004, pp. 135-144.*

Yamada & Soga-03

➙ *K. Yamada and M. Soga, "A compact integrated visual motion sensor for ITS applications", IEEE Trans. On Intelligent Transportation Systems, Vol. 4, No. 1, March 2003, pp. 35-42.*

Zabih & Woodfill-94

➙ *R. Zabih and J. Woodfill, "Non-Parametric Local Transforms For Computing Visual Correspondance", in 3rd European Conference on Computer Vision, 1994, pp 151-158.*

Zahnd et al.-03

➙ *S. Zahnd, P. Lichisteiner, T. Delbruck, "Integrated vision sensor for detecting boundary crossings", Int. Symp. on Circuits and Systems, Vol. 2, 25-28 May, 2003, pp.376-379.*

Zhu et al.-95

➙ *Q. Zhu, T. Lule, H. Stiebig, T. Martin, J. Giehl, J. Zhou, H. Fischer, M. Bohm, "Color array in TFA technology", 4th Int. Conf. On Solid-State and Integrated Circuit Technology, 1995, 24-28 Oct. 1995 pp. 727-729.*

BREVET ET PUBLICATIONS

BREVET

Brevet n°0501630, déposé le 17 février 2005 et extension internationale déposée le 14 février 2006, « Procédé de capture d'images comprenant une mesure de mouvements locaux ».

➙ *L'invention concerne un procédé de capture d'une séquence d'images vidéo, au moyen d'un imageur incluant une estimation des paramètres d'un modèle de mouvement global entre des images successives. Selon l'invention, le procédé comprend une mesure de mouvements locaux sur les bords des images, et l'estimation des paramètres du modèle de mouvement global est réalisée en utilisant le résultat de la mesure de mouvements locaux sur les bords des images. Application notamment à la stabilisation de séquences vidéo.*

CONFERENCES AVEC ACTES ET COMITE DE LECTURE INTERNATIONAL

ESSCIRC'03

D. Navarro, G. Cathebras, and F. Gensolen, "A block matching approach for movement estimation in a CMOS retina: principle and results", in Proc. of the Int. IEEE Conf. on European Solid-State Circuits Conference, Sept. 16-18, 2003, Estoril, Portugal, pp. 615-618

VLSI-SOC'05

F. Gensolen, G. Cathebras, L. Martin, M. Robert, "An integrated image motion sensor for micro camera module", IFIP Int. Conf. On Very Large Scale Integration, Oct. 17-19, 2005, Perth, Australia, pp. 333-338.

ACIVS'05

F. Gensolen, G. Cathebras, L. Martin, M. Robert, "An image sensor with global motion estimation for micro camera module", in Proc. of the Int. IEEE conf. on Advanced Concepts for Intelligent Vision Systems, Sept. 20-23, 2005, Antwerp, Belgium, pp. 713-721.

DDECS'05

F. Gensolen, G. Cathebras, L. Martin, M. Robert, "Pixel level silicon integration of motion estimation", in Proc. of the Int. IEEE workshop on Design and Diagnostic of Electronic Circuits and Systems, April 13-16, 2005, Sopron, Hungary, pp. 93-98.

PRIME'05

F. Gensolen, G. Cathebras, L. Martin, M. Robert, "Focal plane integration of image texture coding for pixel correspondence", in Proc. of the Int. IEEE PhD Research in Microelectronics and Electronics, July 25-28, 2005, Lausanne, Switzerland, pp. 327-330.

CONFERENCES AVEC ACTES ET COMITE DE LECTURE NATIONAL

READ'05

F. Gensolen, G. Cathebras, L. Martin, M. Robert, "Estimation du mouvement global par mesures locales périphériques", Colloque Rétines Electroniques, Asic-FPGA et DSP pour la vision et le traitement d'images en temps réel, 1-3 Juin 2005, INT Evry, France, p. 93-98.

SAME'05

F. Gensolen, G. Cathebras, L. Martin, M. Robert, " An image sensor with global motion estimation for micro camera module", in Proc. of Sophia Antipolis MicroElectronics, October 5-6, 2005, Sophia Antipolis, France.

DIVERS

DOCTISS'05

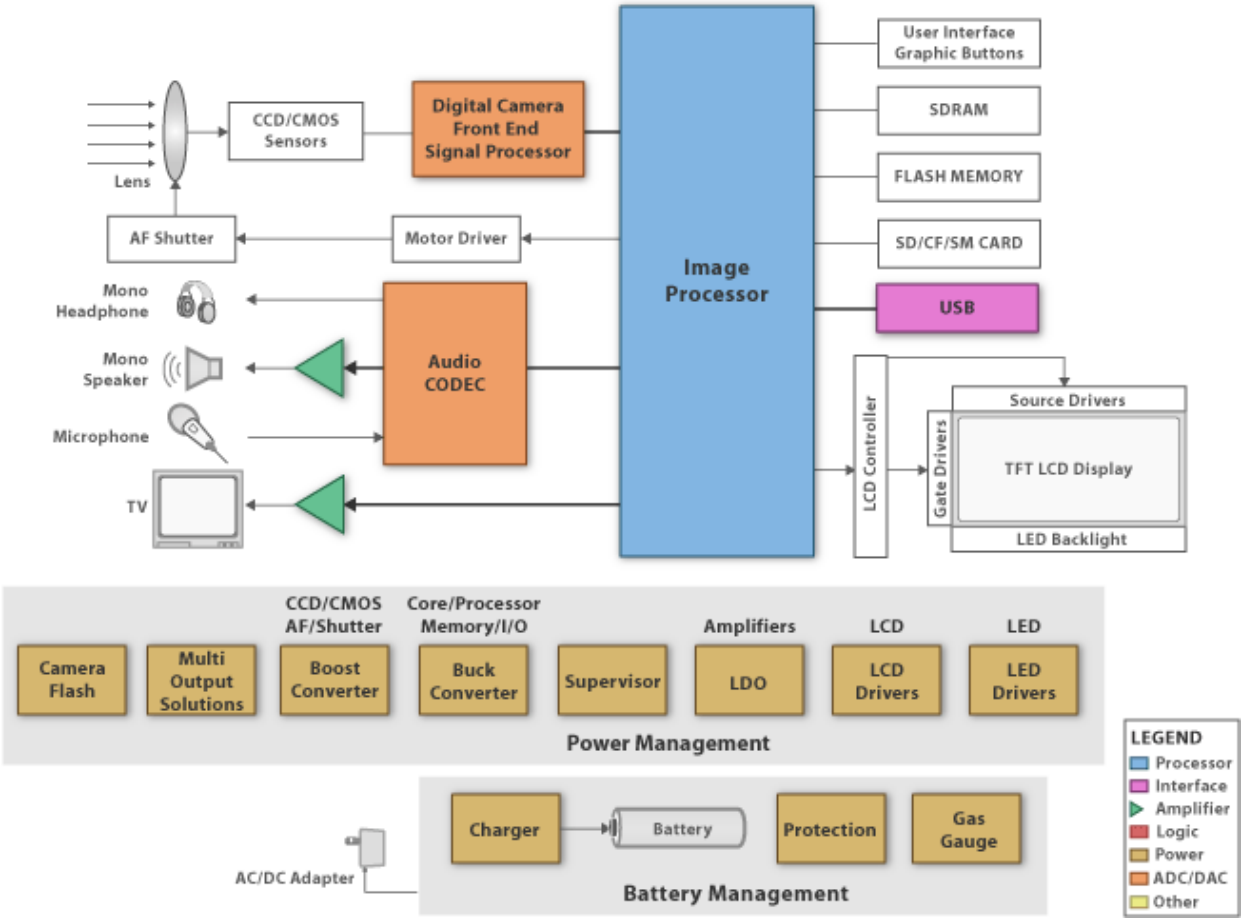
F. Gensolen, G. Cathebras, L. Martin, M. Robert, "Intégration silicium de l'estimation du mouvement", Journée de l'école DOCTorale Information Structure Systèmes, 9 Mars 2005, Montpellier, France.

JNRDM'04

F. Gensolen, G. Cathebras, L. Martin, M. Robert, "Estimation du mouvement global en vue de son intégration sur silicium", Journées Nationales du Réseau Doctoral en Microélectronique, 5-7 Mai 2004, Marseille, France, p. 38-40.

ANNEXES

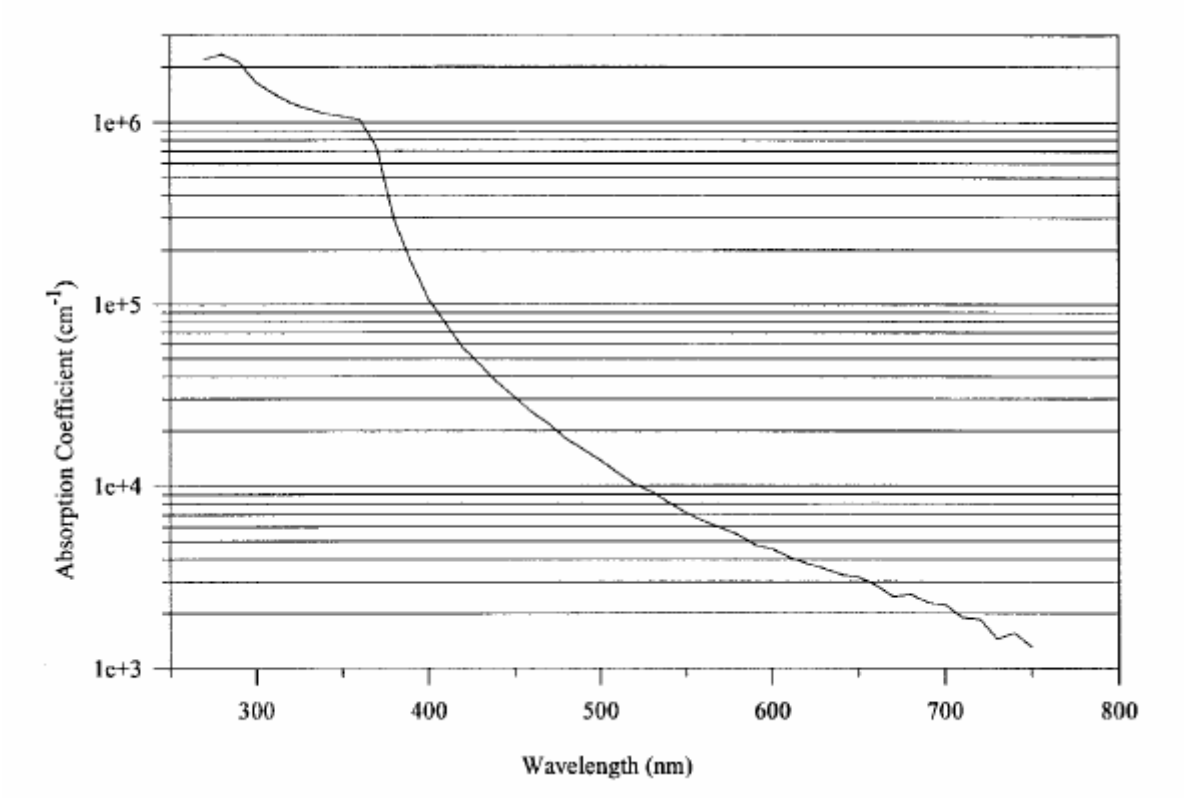
ARCHITECTURE COMPLETE D'UN DISPOSITIF D'ACQUISITION D'IMAGES NUMERIQUES PORTABLE



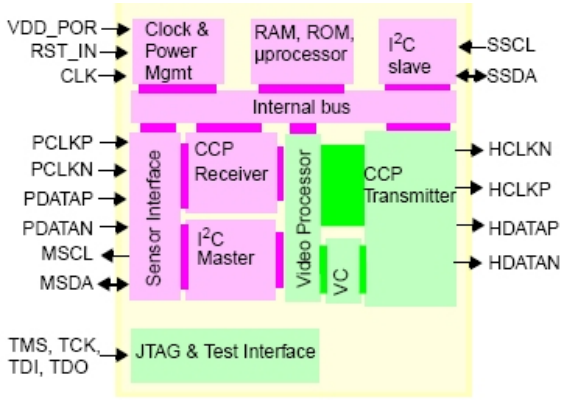
www.ti.com

COEFFICIENT D'ABSORPTION DANS LE SILICIUM

(TEMPERATURE : 300K)

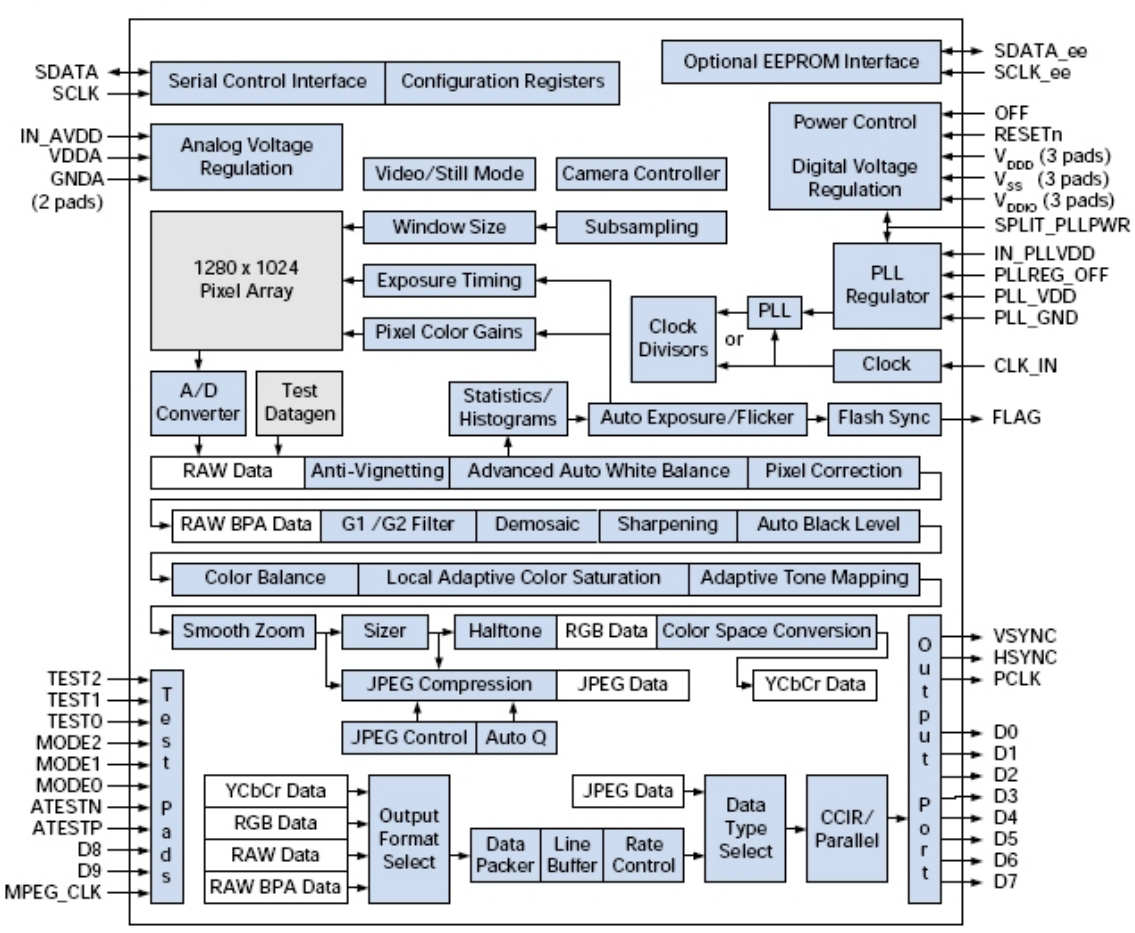


ARCHITECTURE SYSTEME SUR PUCE D'UN IMAGEUR



STV0976, STMicroelectronics, Octobre 2004.

http://www.st.com/stonline/products/applications/consumer/cmos_imaging/



Coprocesseur vidéo associé (ADCC-3960, Agilent, Octobre 2005)

www.agilent.com/semiconductors

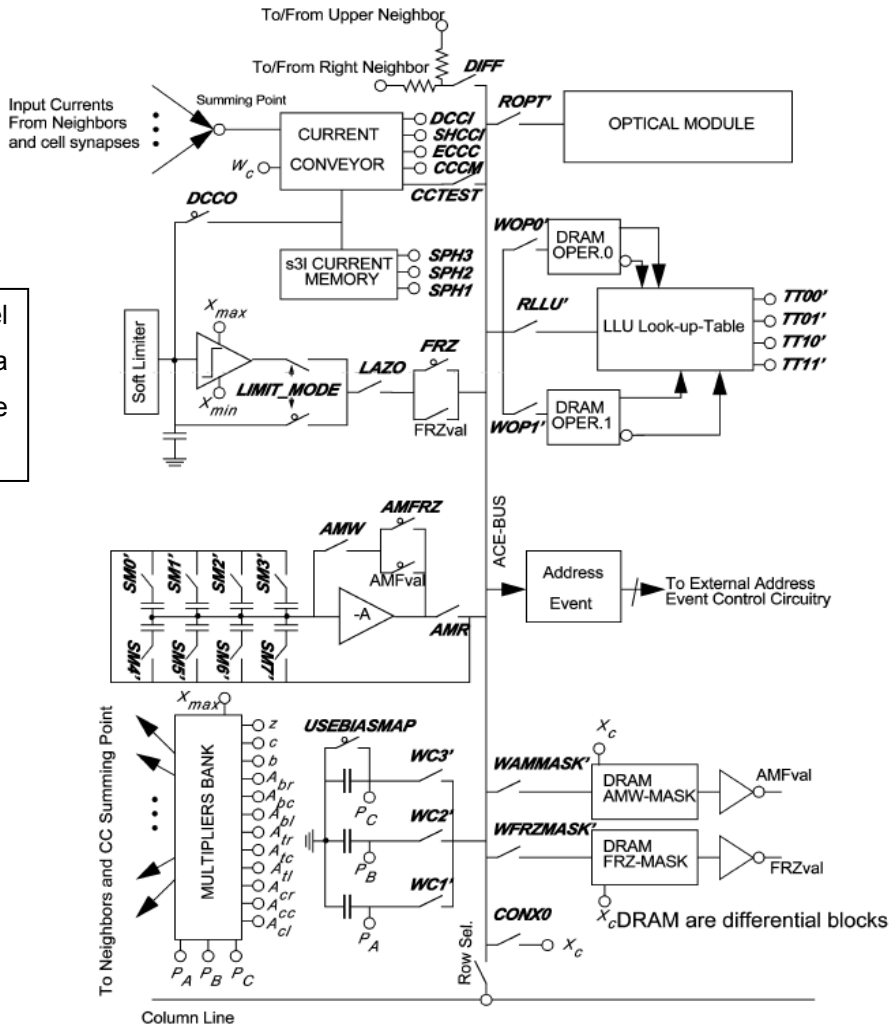
CARACTERISTIQUES DE CIRCUITS DE VISION PROGRAMMBALES

[Rodriguez-Vazquez et al.-04]

Feature	ACE400	ACE4k	ACE16k
Technology @ Supply	0.8µm 2M-1P @ 5V	0.5µm 3M-1P @ 3.3V	0.35µm 5M-1P @ 3.3V
Design Style	Full Custom	Full Custom	Full Custom / Standard Cells
Signal Range	2V (Fully-Diff.)	0.6V – 1.4V	0.6V – 1.4V
Weight Range	2V (Fully-Diff.)	2. 15V – 2.95V	2. 15V – 2.95V
Analog Accuracy	7-bit	7.7-bit	8-bit
# Analog Instructions	8	32	32
# Digital Instructions	N/A	64	4096
# Memories per PE	4 BW	4 BW & 4 Grey	2 BW & 8 Grey
I/O Digital Speed	22 x 10 Mbit/s	16 x 10Mbit/s	120MByte/s
I/O Analog Speed	N/A	16 x 1MSamp/s	
Array Size	22 x 20	64 x 64	128 x 128
# Transistors on Chip	~200.000	~1,000,000	~3.75 mill.
PE Density (PE/mm^2)	27.5	82	180
Computing Power (GOPS)[a]	15.8	40	330
Speed / Area (GOPS/mm^2)	0.98	1	3.2
Speed / Power (GOP/ Joule)	25	39.5	100

a. For the ACE400 chip, speed figure refers to Boolean Operations per Second (BOPS) whereas for ACE4k and ACE16k it refers to 8-bit equivalent resolution operation, i.e. 8-bit additions or products.

Schéma fonctionnel d'une cellule de la matrice du circuit de vision « ACE-16k ».



[Dudek & Hicks-05]

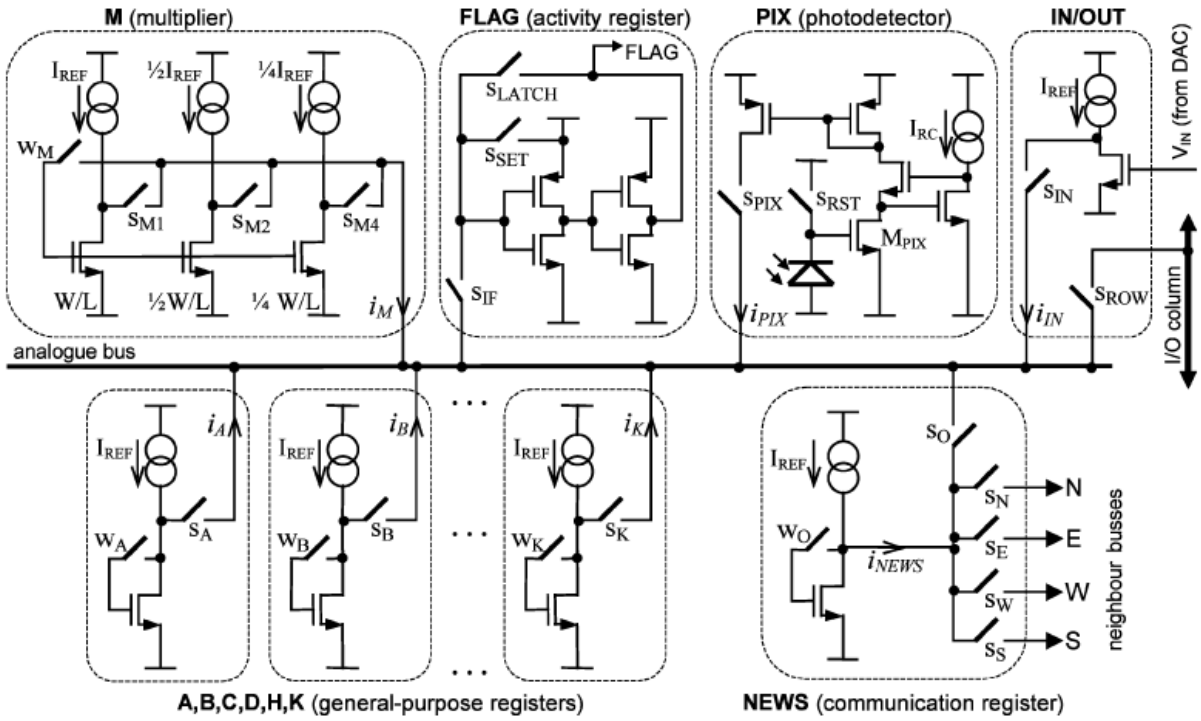


TABLE I
TIME OF EXECUTION OF SEVERAL ALGORITHMS ON THE SCAMP CHIP (NOT INCLUDING READ-OUT TIME)

algorithm	execution time
Smooth using 3×3 convolution template	5.6 µs
Sharpen using 3×3 convolution template	6.0 µs
Edge detection with Sobel templates	11.6 µs
Median Filter in 3×3 neighbourhood	61.6 µs
Binary Morphology (erosion, dilation)	8.0 µs
Conway's Game of Life (per generation)	13.2 µs
Adaptive threshold (threshold level set based on average value in 6×6 neighbourhood)	31.6 µs
Histogram with 64 bins	205.6 µs
Motion estimation (21×21 global block search matching in horizontal direction, with max. displacement ±3 pixels)	46.4 µs
A/D converter (5-bit conversion, ramp)	130.8 µs
D/A converter (5-bit conversion)	11.2 µs

TABLE II
MAIN PARAMETERS OF THE SCAMP CHIP

Technology	0.6 µm CMOS
Array Size	21 × 21
Clock frequency	2.5 MHz
Peak Performance	1.1 GIPS
Pixel pitch	98.6µm
Photodetector fill factor	8.4 %
PE density	102 cells/mm^2
Power per PE (maximum)	85 µW
No. of transistors per PE	128
Memory per PE	8 (analogue)
Error (single transfer)	≈ 0.6%
Accuracy (image filtering)	≈ 2.5%

EXTRAIT DE LA SEQUENCE « GUZET »

EXTRAIT DE LA SEQUENCE « NATURE »

EXTRAIT DE LA SEQUENCE « BUREAU »

ARCHITECTURE ET CONCEPTION DE RETINES SILICIUM CMOS : INTEGRATION DE LA MESURE DU MOUVEMENT GLOBAL DANS UN IMAGEUR.

RESUME : Les capteurs d'images CMOS n'étaient envisagés au début des années 90s que dans le cadre de recherches. La technologie CCD dominait alors. Puis l'évolution extraordinaire des procédés de fabrication de circuits intégrés CMOS a fait qu'aujourd'hui nous avons atteint une égalité en termes de parts du marché. Cette forte croissance est étroitement liée à l'avènement des dispositifs portables grand public tels que les téléphones mobiles, qui embarquent pour la majorité les fonctions photo ou vidéo. En effet, les contraintes d'intégration et de coût favorisent la technologie CMOS. Cependant la prise de vue à l'aide de ces dispositifs portables, très sujets aux tremblements, nécessite une stabilisation de la vidéo qui implique d'estimer le mouvement global inter images. Aussi, l'objectif de ce travail est d'intégrer cette fonction aux imageurs fabriqués par la société STMicroelectronics.
Pour ce faire, une technique novatrice pour estimer ce mouvement global est présentée dans ce mémoire. Cette méthode consiste à extraire un modèle du mouvement global à partir de mesures de déplacements locaux en périphérie des images. Elle a tout d'abord été validée de façon algorithmique, avant d'être intégrée sur silicium. L'architecture finale du capteur se caractérise par une zone photosensible partitionnée en une zone centrale et une zone périphérique. La chaîne de traitement du signal comporte quant à elle un traitement au niveau pixel afin de mesurer les mouvements locaux périphériques. Elle comprend aussi un post-traitement dédié aux tâches d'estimation du modèle du mouvement global ainsi qu'à la compensation du mouvement indésiré.

MOTS-CLES : Capteurs d'images, rétines, CMOS, estimation du mouvement, stabilisation vidéo

ARCHITECTURE AND DESIGN OF CMOS SILICON RETINAS : GLOBAL MOTION SENSING INTEGRATION IN AN IMAGE SENSOR.

ABSTRACT : In the early 90's, image sensors were dominated by CCD technology and CMOS sensors were only developed in research labs. Then a balance within the market shares has been reached thanks to the huge improvement in CMOS integrated circuits fabrication process. This is closely related to the emergence of portable devices like mobile phones, which most often embed photo or video functions. Indeed, such system integration in addition to cost constraints have favored CMOS technology. Nevertheless, video shots with these portable devices, very shaking prone, require a video stabilization which needs to estimate the inter frame global motion of sequences. Therefore, the goal of this work is to add this function to the imagers fabricated by STMicroeletronics.
To do so, a new global motion estimation technique is presented in this thesis. This method consists in extracting a global motion model from the local movements perceived in the periphery of images. This has been first validated by an algorithmic approach, before being integrated on silicon. The final architecture of the sensor has a photosensitive area divided into a central area and a peripheral one. The signal processing chain contains a pixel level processing to measure the peripheral local motions. It includes also a post-processing dedicated to the global motion model estimation and to the unwanted motion compensation.

KEY WORDS : Image sensors, retinas, CMOS, motion estimation, video stabilization

Université de Montpellier II : Sciences et Techniques du Languedoc
LIRMM : Laboratoire d'Informatique, de Robotique, et de Microélectronique de Montpellier
161 rue Ada - 34392 – Montpellier Cedex 5