Descripteurs 2D et 2D+t de points d'intérêt pour des appariements robustes

Manuel Grand-Brochier

N° d'ordre : D.U : 2179
EDSPIC : 538

Université Blaise Pascal - Clermont II

Ecole Doctorale
Sciences Pour L'Ingénieur de Clermont-Ferrand

Thèse

présentée par :

Manuel Grand-brochier

pour obtenir le grade de

Docteur d'Université

Spécialité : Vision pour la robotique

Descripteurs 2D et 2D+t de points d'intérêt pour des appariements robustes.

Soutenue publiquement le 18/11/2011 devant le jury :

Mme Catherine Achard	MCF Univ. Pierre et Marie Curie	Rapporteur
M. Frédéric Jurie	Pr Univ. de Caen	Rapporteur
M. Michel Devy	DR LAAS-CNRS	Examinateur
M. Thierry Chateau	MCF Univ. Blaise Pascal	Examinateur
M. Christophe Tilmant	MCF Univ. Blaise Pascal	Examinateur
M. Michel Dhome	DR LASMEA-CNRS	Directeur de thèse

Résumé

De nos jours les méthodes de vision par ordinateur sont utilisées dans de nombreuses applications telles que la vidéo-surveillance, l'aide à la conduite ou la reconstruction 3D par exemple. Ces différentes applications s'appuient généralement sur des procédés de reconnaissance de formes ou de suivi. Pour ce faire, l'image est analysée afin d'en extraire des amers ou des primitives (contours, fonctions d'intensité ou modèles morphologiques). Les méthodes les plus courantes s'appuient sur l'utilisation de points d'intérêt représentant une discontinuité des niveaux de gris caractérisant un coin dans une image. Afin de mettre en correspondance un ensemble de points d'une image à une autre, une description locale est utilisée. Elle permet d'extraire l'information du voisinage de chaque point (valeurs des pixels, des intensités lumineuses, des gradients). Dans le cas d'applications telles que la vidéo-surveillance ou les caméras embarquées, l'ajout d'une information temporelle est fortement recommandé. Cette généralisation est utilisée au sein du laboratoire pour des projets de type véhicules intelligents (CyCab : véhicule intelligent, VELAC : VEhicule du Lasmea pour l'Aide à la Conduite).

Les travaux de recherche présentés dans ce mémoire ont pour objectif de mettre en œuvre différents outils de détection, description et mise en correspondance de points d'intérêt. Un certain nombre de contraintes a été établi, notamment l'utilisation d'images en niveaux de gris, la robustesse et l'aspect générique de la méthode.

Dans un premier temps, nous proposons une analyse bibliographique des méthodes existantes. Cette dernière permet en effet d'en déduire les paramètres de mise en œuvre ainsi que les principaux avantages et inconvénients.

Nous détaillons par la suite la méthode proposée. La détection des primitives repose sur l'utilisation du détecteur fast-hessien que nous optimisons. L'utilisation d'une description locale basée sur des histogrammes de gradients orientés (HOG) est très répandue et procure de très bons résultats. Nous proposons de compléter son utilisation par un recalage et une mise à l'échelle d'un masque d'analyse elliptique créant ainsi une nouvelle forme de description locale (E-HOG). La mise en correspondance des points d'intérêt se base quant à elle sur une approche par corrélation à laquelle nous ajoutons un coefficient de sélection ainsi qu'une étape de suppression des doublons. Les différents résultats validant notre approche s'appuient sur l'utilisation de transformations synthétiques (vérité terrain) ou réelles.

Nous proposons également une généralisation de notre approche au domaine spatio-temporel, permettant ainsi d'élargir son domaine d'utilisation. Le masque d'analyse précédemment cité est modifié et s'appuie donc sur l'utilisation d'ellipsoïdes. Les tests de validation reposent d'une part sur des séquences vidéo ayant subi des transformations synthétiques et d'autre part sur des séquences réelles issues de la plate-forme PAVIN (Plate-forme d'Auvergne pour Véhicules Intelligents).

Table des matières

Table des figures

Liste des tableaux

Notations

Les notations détaillées ci-dessous ne regroupent qu'une partie des éléments utilisés. Des descriptifs supplémentaires seront donc ajoutés au fur et à mesure, afin de détailler au mieux chaque expression ou résultat mentionné.

- Concernant la notation au sein même d'une image ou d'une séquence vidéo :
  - D'un point de vue continu, une fonction d'intensité est définie par :

$$\begin{aligned} I :&\Gamma \subset \mathbb{R}^d \to \mathbb{R} \\ &\mathbf{x} \mapsto I(\mathbf{x}), \end{aligned} \tag{1}$$

    où $\mathbf{x} = (x, y)$ et $d = 2$ dans le cas d'une image, $\mathbf{x} = (x, y, t)$ et $d = 3$ dans le cas d'une séquence.

  - D'un point de vue discret, une fonction d'intensité se note :

$$\begin{aligned} I :&\Gamma \subset \mathbb{Z}^d \to \mathbb{Z} \\ &\mathbf{m} \mapsto I[\mathbf{m}], \end{aligned} \tag{2}$$

    où $\mathbf{m} = (x, y)^t$ et $d = 2$ dans le cas d'une image, $\mathbf{m} = (x, y, t)^t$ et $d = 3$ dans le cas d'une séquence.

  - Le voisinage d'un point $\mathbf{x}$ se note :

$$\mathrm{V}_I(\mathbf{x}; \Omega) = I(\mathbf{x}) \times \mathbb{1}_\Omega(\mathbf{x}), \tag{3}$$

    où $\Omega \subset \Gamma$ représente une sous-région de l'image et $\mathbb{1}_\Omega$ caractérise la fonction indicatrice ($\mathbb{1}_\Omega(\mathbf{x}) = \begin{cases} 1 \text{ si } \mathbf{x} \in \Omega \\ 0 \text{ sinon} \end{cases}$).

  - Les dérivées partielles et dérivées partielles secondes sont notées respectivement I_a et I_{ab}, avec a et b définis par les variables x, y ou t.
- Concernant les notations et opérations mathématiques :
  - Le signe $*$ correspond au produit de convolution :

$$(f * g)(x) = \int_{-\infty}^{+\infty} f(x - t)g(t)dt = \int_{-\infty}^{+\infty} f(t)g(x - t)dt \tag{4}$$

où f et $g \in L^2(\mathbb{R})$ ou $L^2(\mathbb{C})$.

– Les matrices sont nommées par une majuscule/gras ($\mathbf{M}$, $\mathbf{A}$, $\mathbf{R}$,...), et les vecteurs par une minuscule/gras ($\mathbf{x}$, $\mathbf{vect}$,...).

1 Introduction

De nos jours, l'imagerie numérique devient de plus en plus présente dans les applications courantes de la vie. Elle permet par exemple de surveiller, de localiser, de reconnaitre, ou encore de diffuser des informations. Les méthodes d'obtention de ces images se multiplient, que ce soit par acquisition (scanners, appareils photo, caméscopes numériques, cartes d'acquisition) ou par création (modélisation, DAO : Dessin Assisté par Ordinateur). Depuis des dizaines d'années les scientifiques cherchent et proposent des procédés afin d'acquérir ou de créer de telles images, de les intépréter, de modifier leur contenu, ou encore d'en extraire l'ensemble des informations nécessaires à diverses applications.
Pour en donner quelques exemples, nous pouvons citer la reconstruction 3D [18][90][101][105] permettant de modéliser, à partir d'un certain nombre d'images, l'environnement entourant la ou les caméras. Pour ce type d'applications, la mise en correspondance "inter-image" de points remarquables est nécessaire. Ces appariements et l'exploitation d'un modèle géométrique de caméra permettent la reconstruction par triangulation des points 3D correspondants. La figure 1.1 illustre une reconstruction 3D d'une place de Prague, pouvant être utilisée pour de la localisation par exemple.

FIG. 1.1 – Reconstruction 3D d'une place de Prague (images extraites de [105]).

Il existe également des procédés de suivi d'objets [2][49][77][115], nécessitant une analyse préliminaire (détection et description) afin de caractériser les points extraits des images. Le suivi (ou *tracking*), présenté en figure 1.2, est défini par l'étude du

déplacement au fil du temps de points d'intérêt, dont l'identification se base sur leur descripteur.

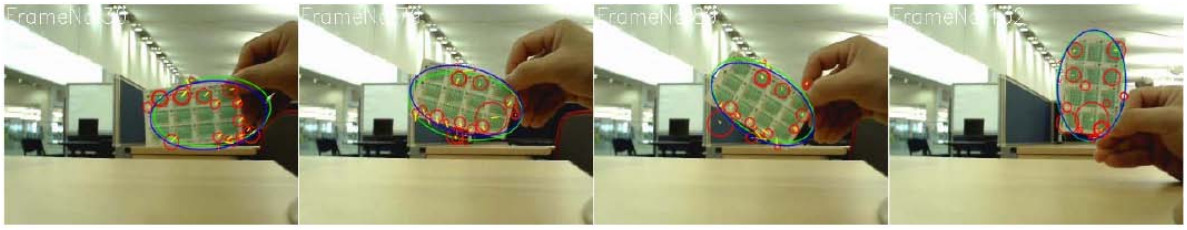


FIG. 1.2 – Suivi d'objets, pour un déplacement entrainant une translation, une rotation et un changement d'échelle (images extraites de [49]).

D'autres procédés tels que le calibrage de caméras [9] ou la stéréo-vision [72][100] sont amenés à utiliser des outils de caractérisation de points et de mise en correspondance. La reconnaissance d'objets [58][66][93] ou de gestes [3][97], s'appuyant également sur ce type d'outil, a la particularité d'utiliser une base d'apprentissage. La comparaison entre cette dernière et l'image analysée permet d'identifier l'objet ou le geste. La figure 1.3 présente deux types de reconnaissance d'objets (voitures et piétons).

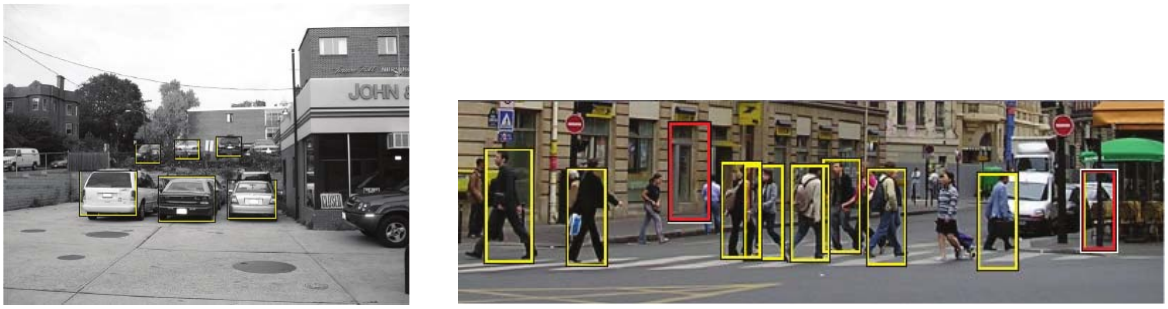


FIG. 1.3 – Reconnaissance d'objets : à gauche des voitures et à droite des piétons (images extraites de [66]).

Les différentes applications citées précédemment s'appuient sur un ensemble de données extraites de l'image. Généralement la première étape consiste à identifier les régions présentant des propriétés locales remarquables (contours, points, textures). Les applications utilisent principalement des points d'intérêt, définissant une double discontinuités de la fonction d'intensité. Les points extraits sont en second lieu caractérisés, le plus souvent localement, afin d'analyser l'information présente dans le masque de description (forme géométrique définissant une région d'intérêt). Une dernière étape permet d'apparier les points d'intérêt présents dans deux images et d'extraire ainsi des couples. La figure 1.4 schématise les différentes étapes de l'analyse pour une mise en correspondance de deux images.

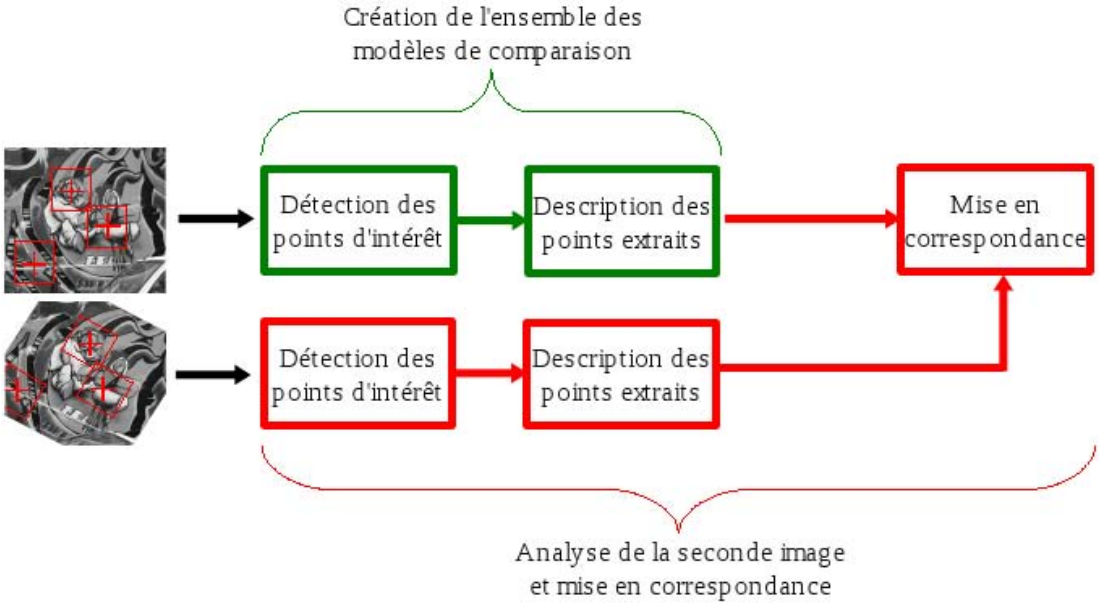


FIG. 1.4 – Etapes de l'analyse d'image et de la mise en correspondance.

La première image permet de créer un ensemble de descripteurs caractérisant les points d'intérêt extraits, qui seront par la suite comparés à l'ensemble des descripteurs issus de la seconde image, afin de réaliser des appariements entre ces points. L'extraction et la caractérisation de points d'intérêt présentent néanmoins certaines limites telles que la répétabilité, la robustesse aux transformations de l'image ou encore les temps de calcul. Ces différentes contraintes ayant des conséquences directes sur les performances des applications citées précédemment, la recherche de nouvelles méthodes ou l'amélioration des approches existantes restent donc très importantes.

1.1 Problématique et objectifs de cette thèse

Les travaux de recherche présentés dans ce manuscrit s'appuient essentiellement sur l'analyse d'images, leurs caractérisations et leurs mises en relation. Nous avons vu précédemment que de nombreuses applications sont susceptibles d'utiliser un tel procédé. Les travaux réalisés dans le cadre de cette thèse ont pour domaine applicatif potentiel les outils algorithmiques liés à la localisation et à la reconstruction de l'environnement. Ces deux procédés dépendant fortement de la qualité des points d'intérêt et de leurs mises en correspondance, il semble important d'accroître les performances de l'analyse faite en amont. Les contraintes à prendre en considération sont principalement liées aux images provenant de la caméra du véhicule. Ces dernières sont en niveau de gris, par conséquent notre méthode se limite à une analyse monospectrale. Les images sont également perturbées par les conditions d'acquisition (météo, route, incident du véhicule), c'est pourquoi la robustesse aux différentes transformations de l'image

(rotations, changements d'échelle, de point de vue, de luminosité) est primordiale. Nous proposons donc d'élaborer une méthode robuste d'extraction de couples de points, axée sur la qualité des appariements. Nous avons pu observer que pour des procédés tels que la localisation, le SLAM (localisation et cartographie en simultané), ou encore la reconstruction 3D, un nombre de points élevé n'est pas nécessaire si la qualité est au rendez-vous. Ainsi nous proposons des techniques favorisant la qualité de la mise en correspondance, au dépend d'une légère diminution du nombre d'appariements.

Il existe un certain nombre de méthodes pouvant être utilisées pour répondre à cette problématique. Dans un souci d'apporter de nouvelles solutions afin d'améliorer l'existant, nous analysons et comparons ces différentes approches. S'appuyant partiellement sur les avantages de chacune d'entre elles, l'objectif est l'élaboration d'un système complet regroupant :

- une méthode d'extraction de primitives dont les données extraites restent cohérentes ;
- une caractérisation robuste du voisinage du point d'intérêt, permettant d'analyser l'information locale et d'en donner une description pertinente ;
- une étape de mise en correspondance, basée sur les différentes données du descripteur et favorisant la qualité des appariements.

Afin de valider cette approche, une comparaison avec des méthodes existantes est nécessaire. Les critères d'observation sont :

- la **précision**, permettant d'évaluer la qualité de la mise en correspondance ;
- le **taux d'appariement** ;
- l'étude de la **robustesse** vis à vis de la détérioration des données.

Au vu des différents objectifs, le système recherché doit donc présenter la meilleure précision possible, tout en conservant un taux d'appariements correct. Il doit également être robuste aux différentes transformations de l'image ainsi qu'aux perturbations des données.

1.2 Solutions proposées

Mes travaux de recherche présentent trois contributions :

- une étude comparative des différentes approches existantes ;
- une nouvelle méthode d'analyse spatiale ;
- une généralisation spatio-temporelle de cette dernière pour des applications utilsant des séquences vidéo.

Souhaitant répondre aux besoins des applications telles que la stéréo-vision ou la reconstruction 3D, nos travaux se limitent tout d'abord au domaine spatial permettant, à partir de deux images, d'en extraire les couples de points avec la meilleure précision possible. L'étude des différentes méthodes existantes nous permet d'en conclure que les approches SIFT [73][74] et SURF [13] sont les plus utilisées et présentent généralement les

meilleurs résultats. En s'appuyant sur les points forts de ces dernières, en les améliorant et en les couplant judicieusement à une exploration adaptative elliptique du voisinage du point d'intérêt, nous proposons la méthode d'analyse spatiale REFA (*Robust E-hog for Features Analysis*) :

- l'étape d'extraction des points d'intérêt repose sur le détecteur fast-hessien ;
- la caractérisation du voisinage se base sur l'utilisation d'histogrammes de gradients orientés, calculés suivant un voisinage elliptique (E-HOG) ;
- la mise en correspondance s'appuie quant à elle sur une méthode classique de minimisation des distances inter-descripteurs.

L'ensemble des modifications, optimisations et outils supplémentaires apportés à notre approche est détaillé dans ce manuscrit. Afin de valider notre méthode, nous proposons une étude comparative entre cette dernière, SIFT et SURF. Une synthèse des différents résultats est proposée, illustrant les performances obtenues pour des transformations diverses, aussi bien synthétiques que réelles.

Les applications utilisant des caméras embarquées, ou le suivi d'objets, s'appuient sur une analyse et une mise en correspondance de séquences vidéo. Notre méthode spatiale ne permettant pas d'obtenir les résultats souhaités pour ce type d'analyse, nous la généralisons au domaine spatio-temporel. Les différentes modifications apportées à notre approche se résument par :

- le détecteur utilisé est le hes-STIP (*hessian spatio-temporal interest point*) ;
- une augmentation du nombre de classes des histogrammes, afin de prendre en compte les données temporelles ;
- la modification de notre masque d'analyse, transformant les ellipses en ellipsoïdes.

Les différentes améliorations et optimisations énoncées précédemment sont conservées. La validation de notre généralisation s'appuie, quant à elle, sur l'étude comparative entre notre système, SIFT3D [99][61] et HOG/HOF [65][63]. Une synthèse est également proposée, regroupant les différents résultats obtenus pour des séquences vidéo de synthèses ou réelles.

Ces différents travaux ont donné lieu aux publications listées page 159.

1.3 Plan du mémoire

Après ce chapitre d'introduction, le mémoire se décompose de la façon suivante :

- le chapitre 2 présente, après une courte définition d'un point d'intérêt, une étude bibliographique des différentes méthodes existantes de détection, de description du voisinage et de mise en correspondance. Cette étude permet notamment de mettre en avant les avantages et inconvénients de chacune d'entre elles, ainsi que leurs étapes de construction ;
- le chapitre 3 regroupe les différentes étapes de construction de notre méthode REFA (choix du détecteur, masque d'analyse, construction des E-HOG, mise en correspondance). Nous y détaillons les différentes améliorations et optimisations

apportées. Nous présentons également une analyse comparative des performances de notre approche, du SIFT et du SURF pour des transformations aussi bien synthétiques que réelles. Afin de valider pleinement notre méthode, une synthèse des résultats obtenus est proposée et des tests sur l'estimation d'homographie apparaissent en fin de chapitre ;

– le chapitre 4 décrit la généralisation de notre approche au domaine spatio-temporel. Afin d'apporter les modifications nécessaires à notre méthode REFA, une étude bibliographique des outils et procédés existants pour ce domaine est proposée. Nous détaillons alors les différents changements apportés à notre approche. La validation de cette généralisation se base sur une analyse comparative avec les méthodes SIFT3D et HOG/HOF. Nous proposons également une synthèse des différents résultats obtenus ainsi que des tests de recalage de sous-séquences ;
– Nous concluons au chapitre 5 par une synthèse des solutions proposées, couplée à une analyse et une critique des différents résultats obtenus. Des perspectives concernant notamment l'amélioration de ces travaux et leurs intégrations aux véhicules intelligents sont également abordées.

2 Etat de l'art

2.1 Définition d'un point d'intérêt

La détection de points d'intérêt, tout comme la détection de contours, est une étape préliminaire à de nombreux processus dans le domaine de la vision par ordinateur. Les points ainsi extraits peuvent être utilisés dans les procédés de reconstruction 3D, de suivi d'objets, de détection de personnes ou encore de reconnaissance de gestes. Les points d'intérêt correspondent généralement à une discontinuité des niveaux de gris comme le montre la figure 2.1. Ils peuvent également apparaître lors d'une modification de la structure, de la texture ou de la géométrie de l'image.

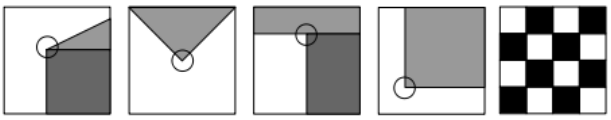


FIG. 2.1 – Différents types de points d'intérêt : coin simple, jonction en 'V', jonction en 'T', jonction en 'L', jonction en 'damier'.

Le choix de l'utilisation d'un détecteur de coins, plutôt qu'un détecteur de contours, est souvent préconisé pour déterminer les points d'intérêt. En effet, du fait de la grande diversité de ce type de détecteurs, il est plus aisé de gérer les différentes problématiques liées à l'image (bruit, transformations, occultations). Néanmoins, nous verrons que pour certains détecteurs de coins, seul un seuil les différencie d'un détecteur de contour. Afin de faciliter ce choix, il est possible de les classer suivant trois méthodes de détection :

1. méthode basée sur les contours. Introduit en 1986 par Asada et Brady [7], puis repris en 1993 par Deriche et Giraudon [34], le procédé consiste à détecter des contours, puis à les seuiller pour en extraire les points d'intérêt.
2. méthode basée sur l'utilisation de modèle morphologique mathématique, introduit en 1995 par Zhang et Zhao [113] puis repris en 2004 par Dinesh et Guru [37].
3. méthode basée sur l'étude de la fonction d'intensité en chaque pixel de l'image. Quelques exemples tels que Moravec [85] en 1977, Harris [47] en 1988, ou encore MSER [1][78] en 2002. Cette méthode reste la plus utilisée, nous proposons donc

[1] *Maximally Stable Extremal Regions*

une étude plus détaillée de ce type de détecteur dans le paragraphe suivant.

2.2 Les détecteurs de points d'intérêt

Le choix d'un détecteur de points d'intérêt repose essentiellement sur l'utilisation souhaitée. Il faut par conséquent les classer afin de faciliter ce choix. L'étude suivante se limite à l'espace 2D, des éléments complémentaires seront apportés ultérieurement afin de gérer l'espace spatio-temporel. La figure 2.2 propose une classification chronologique des différents détecteurs étudiés.

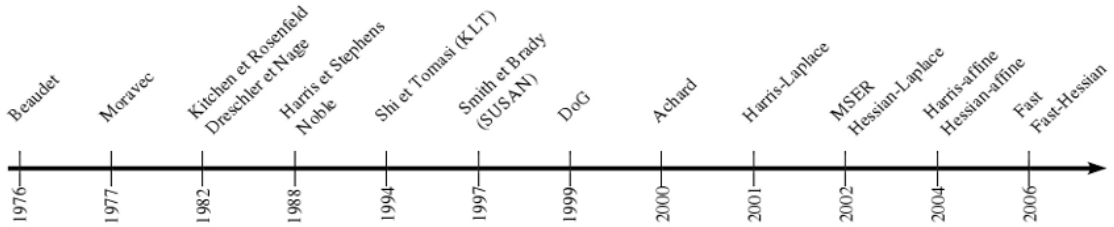


FIG. 2.2 – Classification chronologique des détecteurs étudiés.

Une classification suivant le type d'invariances gérées peut également s'avérer judicieux. Il sera en effet plus aisé de sélectionner le détecteur approprié afin de pallier aux différentes transformations de l'image. Elles permettent notamment de modéliser des changements de point de vue intra caméra définissant un déplacement de cette dernière (rotations, translations) et inter caméra mettant en jeu des réseaux de capteurs (stéréo-vision). Les transformations étudiées sont réparties en quatres catégories.

– Les **transformations euclidiennes** (ou rigides) se composent de l'identité, de la rotation et de la translation. Elles préservent les angles, les distances et sont inversibles. La figure 2.3 donne un aperçu des modifications de l'image obtenues par le biais de ces transformations.

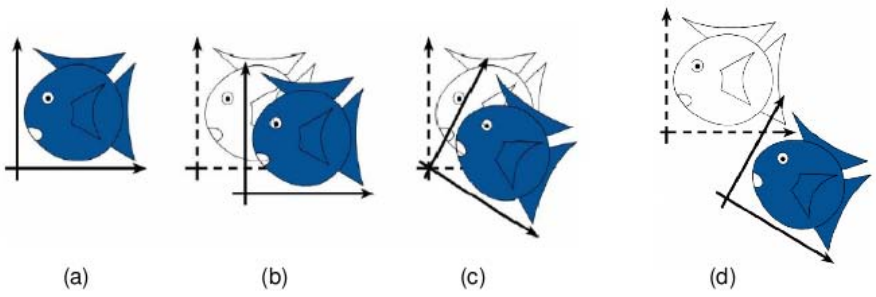


FIG. 2.3 – Transformations géométriques entre le repère 1 (gras) et le repère 2 (pointillé) dans le cas rigide : (a) identité, (b) translation, (c) rotation, (d) exemple de transformation euclidienne.

– L'ajout du changement d'échelle isotrope aux précédentes transformations, permet d'obtenir les **similitudes** (figure 2.4). Ces dernières préservent les angles, le rapport des longueurs et sont inversibles.

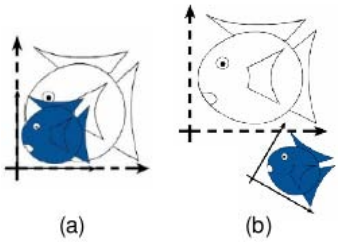


FIG. 2.4 – Composantes des transformations de type similitude entre deux repères (gras et pointillé) : (a)changement d'échelle isotrope, (b)exemple de similitude.

– La troisième catégorie correspond aux **transformations affines**. Ces dernières englobent les deux premiers types de modifications de l'image auxquels s'ajoutent, la réflection, le changement d'échelle anisotrope et le '*shear*[1]'. Cette catégorie conserve les parallèles et est inversible également. La figure 2.5 représente les différentes transformations ajoutées, et la résultante qui en découle.

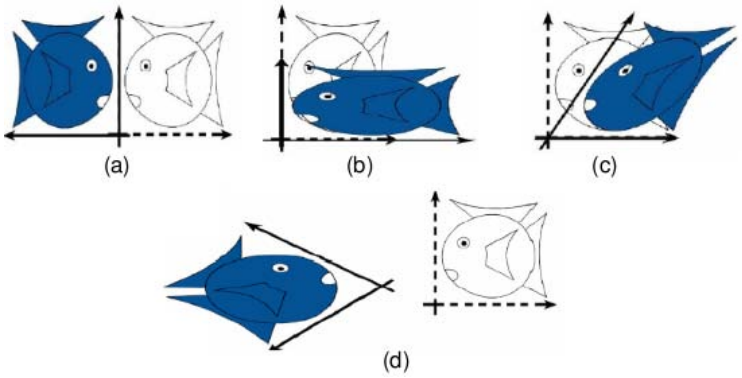


FIG. 2.5 – Transformations affines entre le l'image 1 (gras) et l'image 2 (pointillée) : a) réflection, (b) changement d'échelle anisotrope (c) 'shear' (d) exemple de transformation affine.

– Une dernière catégorie, nommée **transformations projectives**, est obtenue en couplant l'ensemble des transformations précédentes avec une modification de la perspective de l'image. Elles préservent les droites mais ne conservent pas le barycentre. La figure 2.6 représente une transformation perspective de l'image.

[1]défini la non-orthogonalité du référentiel image

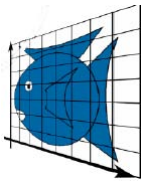


FIG. 2.6 – Transformation projective : exemple de perspective.

Les trois catégories retenues pour la classification des détecteurs peuvent donc être synthétisées par un diagramme (figure 2.7) représentant la répartition des différentes modifications apportées à l'image.

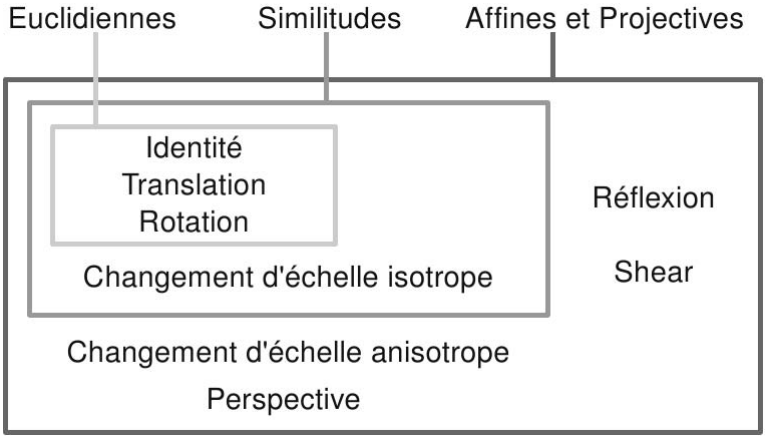


FIG. 2.7 – Schéma des différents types de transformations étudiées.

Une précision concernant les transformées affines et projectives, nous les rassemblons en s'appuyant sur le fait qu'aucun élément supplémentaire n'est apporté aux méthodes affines pour pallier aux problèmes de perspectives.

2.2.1 Invariances aux transformations euclidiennes

Dès 1976, de nombreuses méthodes de détection de points d'intérêt ont vu le jour. Certaines d'entre elles se basent sur l'utilisation d'une matrice hessienne [15] [39], [60], d'autres s'appuient sur l'analyse du changement local d'intensité [1] [47] [85] [94] [102] [89] [104]. Une description de ces différentes méthodes est donc nécessaire afin d'en détailler la construction et de définir les relations existantes entre elles.

2.2.1.1 Détecteurs : Beaudet, Dreschler-Nagel, Kitchen-Rosenfeld

Nous allons étudier des détecteurs se basant sur l'utilisation des dérivées partielles secondes, comme les détecteurs de Beaudet [15], de Dreschler-Nagel [39] ou encore de Kitchen-Rosenfeld [60]. D'un point de vue théorique, une analyse s'appuyant sur la matrice hessienne, permet de déterminer la nature des points critiques d'une fonction.

Soit une fonction f à n variables, notées $(x_i)_{1\leq i\leq n}$, dont toutes les dérivées partielles secondes existent, la matrice hessienne se note :

$$\mathbf{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} = \begin{bmatrix} f_{x_1^2} & f_{x_1 x_2} & \cdots & f_{x_1 x_n} \\ f_{x_2 x_1} & f_{x_2^2} & \cdots & f_{x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{x_n x_1} & f_{x_n x_2} & \cdots & f_{x_n^2} \end{bmatrix} \tag{2.1}$$

On définit un point critique pouvant être, soit dégénéré lorsque le hessien (déterminant de la matrice hessienne) s'annule, soit non dégénéré et dans ce cas, il faut étudier sa nature (point d'extremum local ou point col) à travers le signe des valeurs propres de la matrice $\mathbf{H}$:

- si ces dernières sont positives, le point constitue un minimum local ;
- si elles sont négatives, il constitue un maximum local ;
- s'il y a des valeurs propres de chaque signe, le point définit un point selle (point col).

Beaudet [15] propose en 1976, en partant de la matrice hessienne $\mathbf{H}$ (équation 2.1) appliquée à la fonction I de l'image, de calculer une métrique k proportionelle au hessien :

$$k(\mathbf{x}) = C\det(\mathbf{H}(g_\sigma * I(\mathbf{x}))) \tag{2.2}$$

où C est une constante positive et où les dérivées secondes partielles sont lissées par une gaussienne g_σ :

$$g_\sigma(\mathbf{x}) = \frac{1}{2\pi\sigma^2} exp^{-\frac{\|\mathbf{x}\|_2^2}{2\sigma^2}} \tag{2.3}$$

L'auteur se concentre sur la recherche des maxima locaux. Chacun d'entre eux est défini par :

$$\mathbf{x}_B = \underset{\mathbf{x}}{\operatorname{argmax}}(|k(\mathbf{x})|) \tag{2.4}$$

Afin de déterminer les coins, deux critères de sélection sont choisis. Tout d'abord les points critiques doivent être non dégénérés, et ensuite les valeurs propres de la matrice $\mathbf{H}$ doivent être toutes du même signe.

Cette méthode d'analyse est reprise par Dreschler et Nagel en 1982 [39], qui apportent notamment une amélioration sur la sélection des points d'intérêt. En partant de l'équation 2.2 et en s'appuyant sur la propriété du changement de signe de la courbure aux abords d'un coin, ils déterminent un maximum $\hat{\mathbf{x}}_1$ puis cherchent localement un minimum $\hat{\mathbf{x}}_2$ de k :

$$\mathbf{x}_1 = \underset{\mathbf{x}}{\operatorname{argmax}}(k(\mathbf{x})) \qquad et \qquad \mathbf{x}_2 = \underset{\mathbf{x}}{\operatorname{argmin}}(k(\mathbf{x})) \tag{2.5}$$

La ligne joignant ces deux extréma, permet de déterminer l'endroit où la pente du signal est maximale, c'est à dire le point d'annulation de la courbure :

$$\mathbf{x}_{DN} = \frac{\mathbf{x}_1 + \mathbf{x}_2}{2} \tag{2.6}$$

Une autre approche, basée sur une approximation locale polynômiale de la fonction I de l'image autour du point $\mathbf{x}_0 = (x_0; y_0)$ a été proposée par Kitchen-Rosenfeld [60] en 1982 puis repris par Zuniga-Haralick [116] en 1983. Tous deux se basent sur le polynôme bi-cubique suivant :

$$I(x,y) \underset{\mathbf{x}_0}{\approx} c_1 + c_2 x + c_3 y + c_4 x^2 + c_5 xy + c_6 y^2 + c_7 x^3 + c_8 x^2 y + c_9 xy^2 + c_{10} y^3 \tag{2.7}$$

Cette équation permet d'obtenir pour le premier :

$$k_{KR}(\mathbf{x}_0) = \frac{-(c_2^2 c_6 - 2c_2 c_3 c_5 + c_3^2 c_4)}{c_2^2 c_3^2} \tag{2.8}$$

et pour le second :

$$k_{ZH}(\mathbf{x}_0) = \frac{-(c_2^2 c_6 - 2c_2 c_3 c_5 + c_3^2 c_4)}{(c_2^2 c_3^2)^{\frac{3}{2}}} = \frac{k_{KR}(\mathbf{x}_0)}{c_2^2 c_3^2} \tag{2.9}$$

La sélection des points d'intérêt se fait de façon identique à celle du détecteur de Beaudet (équation 2.4). La même année, Kitchen-Rosenfeld proposent une seconde approche, se basant sur l'étude des variations des gradients le long des contours. En s'appuyant sur l'équation 2.8, ils proposent le taux de variation suivant :

$$k'_{KR}(\mathbf{x}_0) = \frac{I_{xx} I_y^2 + I_{yy} I_x^2 - 2 I_{xy} I_x I_y}{I_x^2 + I_y^2} \tag{2.10}$$

Le score ainsi obtenu permet de détecter les zones à forte modification de gradient suivant les différents axes étudiés et d'en prélever les points d'intérêt. En conclusion à ce dernier détecteur, Nagel [88] montre que son approche et celle de Kitchen-Rosenfeld sont identiques.

2.2.1.2 Détecteurs : Harris, Noble, KLT, Achard

Une autre approche, basée sur l'utilisation des dérivées premières partielles, peut être utilisée (détecteurs de Harris [47], Noble [89], KLT [102], et Achard [1]). Cette méthode de détection de points d'intérêt consiste à observer les changements locaux de l'intensité caractérisés par la fonction E :

$$E(\mathbf{x}) = \sum_{\mathbf{k} \in \Omega} \mathbb{1}_\Omega(\mathbf{k}) (I(\mathbf{k} + \mathbf{x}) - I(\mathbf{k}))^2, \tag{2.11}$$

où $\mathbf{x}$ caractérise le vecteur déplacement. Cette équation permet, entre autre, d'obtenir l'orientation et la norme des gradients, pour chaque pixel de l'image. D'un point de vue mathématique, le gradient d'une fonction $f(x,y)$ est un vecteur noté $\overrightarrow{\nabla}$:

$$\overrightarrow{\nabla f} = [f_x, f_y]^T \tag{2.12}$$

Le vecteur gradient pointe dans la direction du plus fort changement d'intensité, et sa longueur dépend de son taux de variation. Une représentation de gradients est donnée en figure 2.8.

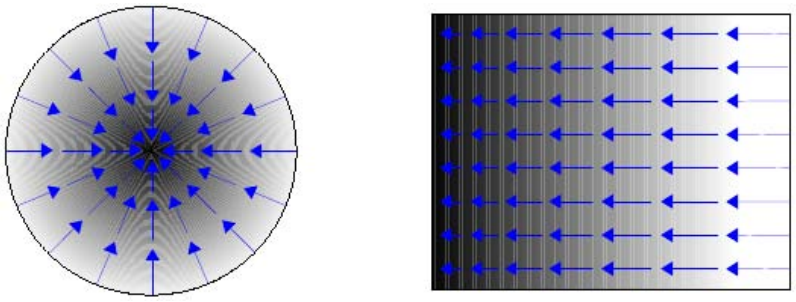


FIG. 2.8 – Différents gradients représentant les changements d'intensité (sombre $\mapsto$ clair).

Dans leur approche faite en 1988, Harris et Stephens [47] proposent de modifier la fonction E de l'équation 2.11 afin de pallier à la faible discrétisation de cette dernière. Ils utilisent par conséquent un développement de Taylor de la fonction d'intensité I autour du point $\mathbf{x}_0 = (x_0; y_0)$:

$$I(x; y) \underset{(x_0;y_0)}{\approx} I(x_0; y_0) + xI_x + yI_y + o(x^2; y^2) \tag{2.13}$$

pour obtenir, après identification :

$$E(\mathbf{x}) = \sum_{\mathbf{k}\in\Omega} \mathbb{1}_\Omega(\mathbf{k})[xI_x + yI_y + o(x^2; y^2)]^2 \tag{2.14}$$

Pour les faibles déplacements, $o(x^2, y^2)$ peut être négligé et finalement la fonction E devient :

$$E(\mathbf{x}) = x^2(I_x^2 \times \mathbb{1}_\Omega)(\mathbf{x}) + 2xy(I_xI_y \times \mathbb{1}_\Omega)(\mathbf{x}) + y^2(I_y^2 \times \mathbb{1}_\Omega)(\mathbf{x}) \tag{2.15}$$

Ils choisissent par la suite de remplacer le filtre binaire par un filtre gaussien g_σ, entrainant un débruitage du signal et donc une amélioration de la réponse du détecteur. L'équation précédente devient :

$$E(\mathbf{x}) = x^2(I_x^2 * g_\sigma)(\mathbf{x}) + 2xy(I_xI_y * g_\sigma)(\mathbf{x}) + y^2(I_y^2 * g_\sigma)(\mathbf{x}) \tag{2.16}$$

Harris et Stephens mettent également en place une méthode de sélection de points d'intérêt basée sur l'analyse du comportement général de E. En effet, il est possible d'en extraire le tenseur de structure noté $\mathbf{M}$:

$$E(\mathbf{x}) = \mathbf{x}\mathbf{M}\mathbf{x}^t \qquad \text{avec} \qquad \mathbf{M} = g_\sigma * \begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} \tag{2.17}$$

Les valeurs propres de $\mathbf{M}$ correspondent aux courbures principales de la fonction E. Le détecteur de Harris se base sur l'hypothèse qu'un point est dit d'intérêt si les valeurs des deux courbures sont élevées. Ceci peut se caractériser par l'analyse des valeurs propres de $\mathbf{M}$, notées λ_1 et λ_2 avec $\lambda_1 \geq \lambda_2$:

– si $\lambda_1 = \lambda_2 = 0$: la zone sélectionnée est complètement uniforme (zone homogène).
– si $\lambda_1 > \lambda_2 = 0$: la zone correspond à un contour et le vecteur propre associé à λ_1 lui est perpendiculaire.
– si $\lambda_1 > \lambda_2 > \varepsilon$ (avec ε étant un seuil) : la zone caractérise un coin.

La figure 2.9 représente l'influence des valeurs propres de M sur le voisinage d'un point.

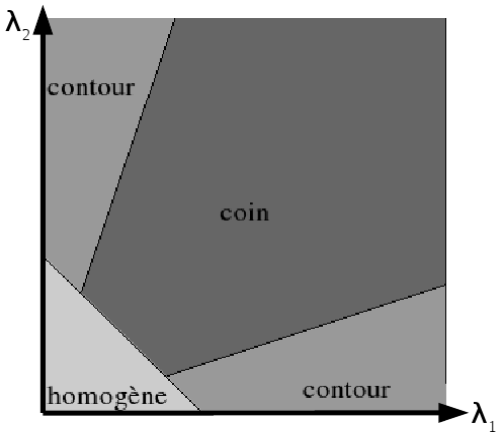


FIG. 2.9 – Schéma simplifié de l'analyse des valeurs propres.

Un point est donc considéré comme point d'intérêt si ses valeurs propres λ_1 et λ_2 sont élevées. Ne souhaitant pas déterminer directement ces dernières, Harris et Stephens utilisent les relations suivantes :

$$\det(\mathbf{M}) = \prod_i \lambda_i \qquad \text{et} \qquad \text{trace}(\mathbf{M}) = \sum_i \lambda_i \tag{2.18}$$

et proposent la mesure k_H suivante :

$$k_H(\mathbf{x}) = \det(\mathbf{M}) - \alpha \text{trace}(\mathbf{M})^2 \tag{2.19}$$

où $\alpha \in [0,02; 0,06]$ est une constante et k_H est communément appelé critère de Harris. La dernière étape de leur détecteur est la recherche de maxima locaux de k_H.

La même année, Noble [89] propose une méthode similaire. La seule différence résulte dans la détermination de son critère défini par :

$$k_N(\mathbf{x}) = \frac{\text{trace}(\mathbf{M})}{\det(\mathbf{M})} \tag{2.20}$$

En 1994, Shi et Tomasi [102] reprennent l'idée d'utiliser la matrice $\mathbf{M}$ pour créer le KLT (Kanade-Lucas-Tomasi feature tracker), mais contrairement au détecteur de Harris, ils choisissent de calculer les valeurs propres. Leur idée est de sélectionner des primitives qui peuvent être facilement suivies, lors d'un déplacement de la caméra par exemple. Afin de détecter les points d'intérêt, deux critères sont donc mis en place :

– les deux valeurs propres doivent être grandes ;
– en partant de $\lambda_1 \geq \lambda_2$ et sachant que la valeur de λ_1 est limitée par l'étendu des valeurs en niveaux de gris, il faut donc émettre une condition sur λ_2 pour respecter le premier critère.

Par conséquent, un point $\mathbf{x}$ est conservé si la valeur propre λ_2 aux coordonnées (x,y) est supérieure à un certain seuil.

En s'appuyant sur les idées du détecteur de Harris, Achard et al. [1] proposent une nouvelle approche basée sur la propriété des angles entre gradients. Cette dernière stipule que dans le cas d'un coin, les angles entre le gradient du point analysé et les gradients de ses plus proches voisins, sont grands. De ce fait pour un point $\mathbf{x}$, ils posent :

$$k_A(\mathbf{x}) = \frac{I_x^2(I_y^2 * m)(\mathbf{x}) + I_y^2(I_x^2 * m)(\mathbf{x}) - 2I_xI_y(I_xI_y * m)(\mathbf{x})}{(I_x^2 * m + I_y^2 * m)(\mathbf{x})} \tag{2.21}$$

où m défini la moyenne sur le voisinage. Pour leur détecteur, Achard et al. ont opté pour un filtre de type :

$$m = \frac{1}{8}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} \tag{2.22}$$

La sélection des points d'intérêt est faite de façon similaire à celle du détecteur de Harris. En effet, l'équation 2.21 peut s'identifier à l'équation 2.15, proposant toutes les deux un critère de sélection. Le seuillage de ce dernier permet donc d'extraire les points représentant un coin.

Nous pouvons apporter quelques détails supplémentaires concernant les détecteurs multi-spectraux. Achard et al. [1] et Gouet et al. [45] proposent deux méthodes s'appuyant notamment sur l'opérateur de Zenzo [112]. Pour la première, l'ajout de données multi-spectrales permet de généraliser l'équation 2.21 et son développement définit k_A comme étant égal à :

$$\begin{aligned} k_A(\mathbf{x}) = & \frac{\alpha(\mathbf{x})}{(m * \alpha(\mathbf{x}))}[m * ((1 - 2cos^2\theta(\mathbf{x}))\lambda cos^2\theta(\mathbf{x}) \\ & + \lambda cos^2\theta(\mathbf{x}) - 2sin\theta(\mathbf{x})cos\theta(\mathbf{x})(\lambda sin\theta(\mathbf{x})cos\theta(\mathbf{x})))] \end{aligned} \tag{2.23}$$

où α correspond au carré de la norme du gradient de I et θ à son orientation. Pour la seconde proposition, Gouet et al. extraient tout d'abord les contours à l'aide de l'opérateur de Zenzo. Les points d'intérêt sont extraits par seuillage de la valeur maximale de la courbure d'intensité.

2.2.1.3 Détecteurs : Moravec, SUSAN, FAST

Dans l'optique de simplifier l'analyse du voisinage et/ou d'optimiser les temps de calculs, certains détecteurs préfèrent baser leur étude uniquement sur la fonction d'intensité de l'image.

Moravec [85] est un des premiers à s'intéresser à la fonction d'intensité I. En 1977, il propose d'utiliser la fonction E (équation 2.11), permettant notamment de déterminer les irrégularités dans un signal, pour créer son détecteur. L'analyse locale qu'il propose se base sur l'interprétation des différentes situations suivantes :

- dans le cas 1, la fonction E a de faibles valeurs dans toutes les directions :
  $\rightarrow$ zone homogène.
- dans le cas 2, la fonction E a de faibles valeurs dans une direction et de fortes valeurs dans la direction normale :
  $\rightarrow$ zone ayant un contour.
- dans le cas 3, la fonction E a de fortes valeurs dans toutes les directions :
  $\rightarrow$ zone ayant un coin.

La sélection des points d'intérêt va donc être focalisée sur ce dernier cas.

$$\underset{\mathbf{x}}{\operatorname{argmin}}(E(\mathbf{x})) > \epsilon \tag{2.24}$$

où ϵ est un seuil, déterminé de façon à sélectionner uniquement les points les plus pertinents. Le principe de ce détecteur est donc de rechercher les maxima locaux de la valeur minimale de E. Il faut préciser qu'en 1988, Harris et Stephens ont voulu, au vu des bons résultats fournis par ce détecteur, reprendre les idées de Moravec, et les améliorer. Ceci, comme il a été décrit en 2.2.1.2, dans le but de le rendre plus stable et plus performant.

En 1997, Smith et Brady [104] proposent un nouveau détecteur, nommé SUSAN (*Smallest Univalue Segment Assimilating Nucleus*), qui se base sur une analyse locale circulaire. La figure 2.10 donne un exemple du masque d'analyse utilisé.

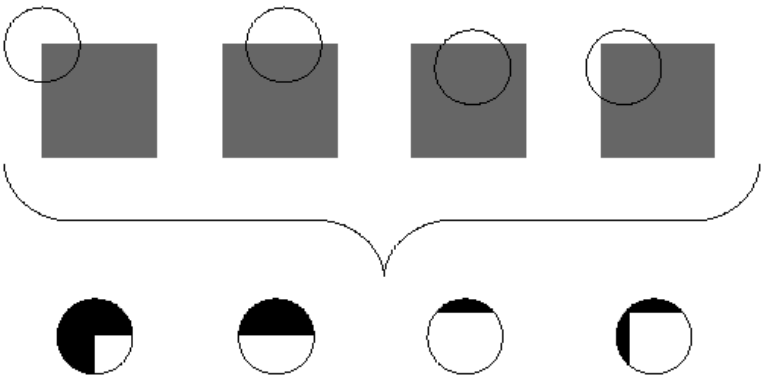


FIG. 2.10 – Exemple de masques d'analyse du détecteur SUSAN (haut), permettant le calcul des zones USAN (bas)

Pour chacun des cas observés ci-dessus, la somme des comparaisons entre chaque pixel du cercle, noté C_Ω, et le niveau de gris de son centre est déterminée par :

$$n(\mathbf{x}_0) = \sum_{\mathbf{d} \in C_\Omega} e^{\frac{(I(\mathbf{x}_0 - \mathbf{d}) - I(\mathbf{x}_0))^6}{\epsilon}} \tag{2.25}$$

où $\mathbf{x}_0$ est le centre du cercle et ϵ correspond à un seuil. Ce calcul permet ainsi de déterminer la zone USAN (*Univalue Segment Assimilating Nucleus*), représentant le nombre de pixels de même niveau de gris que $\mathbf{x}_0$ (figure 2.10). La réponse du détecteur est donnée par :

$$R_S(\mathbf{x}_0) = \begin{cases} \frac{\max(n)}{2} - n(\mathbf{x}_0) & \text{si } n(\mathbf{x}_0) < \frac{\max(n)}{2} \\ 0 & \text{sinon} \end{cases}, \tag{2.26}$$

ce qui revient à dire que si le centre du cercle d'analyse se trouve sur un coin, alors la taille de la zone USAN sera strictement plus petite que la demi-surface du masque.

Ayant pour objectif la diminution des temps de calculs, Rosten et Drummond [94] mettent en place une nouvelle approche de sélection de points d'intérêt, le détecteur FAST. Se basant sur les deux précédentes méthodes, ils proposent une analyse circulaire locale de chaque pixel. La figure 2.11 donne une vue d'ensemble de l'analyse du pixel p et du voisinage concerné (pixels 1 à 16).

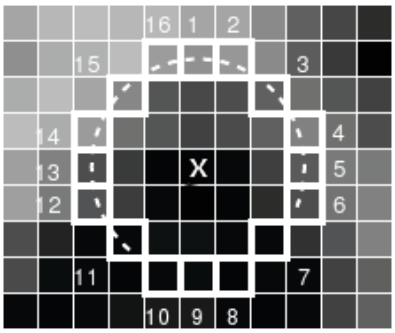


FIG. 2.11 – Représentation de l'analyse du détecteur FAST (image extraite de [94]).

Dans un premier temps, la valeur $I(\mathbf{x})$ est comparée avec celle des pixels correspondant aux points cardinaux (pixels 1,5,9 et 13). Cela permet d'effectuer une sélection préliminaire, en s'appuyant sur le fait qu'un point $\mathbf{x}$ est conservé si au moins trois des quatres pixels de comparaison sont, soit plus clairs, soit plus sombres que lui. Le voisinage des points retenus est analysé suivant la figure 2.11 puis classé en trois catégories (sombre, homogène ou claire). Cette classification permet d'extraire rapidement les points d'intérêt de l'image. En effet, dans le cas où douze des seize pixels sont consécutivement dans la même catégorie, $\mathbf{x}$ est considéré comme étant un coin.

2.2.2 Invariances aux similitudes

Nous avons vu dans la figure 2.7 que les similitudes sont l'ajout du changement d'échelle isotrope aux transformations étudiées précédemment. Afin d'y être le plus robuste possible, notre étude s'est portée sur l'utilisation des espaces d'échelles. Introduits en 1984 par Koenderink [62], ils permettent notamment, de gérer les différents

changements d'échelles subis par l'image. Ils sont définis comme étant la représentation multi-résolution d'un signal. En effet, un espace d'échelles permet, en partant de l'image d'origine, de supprimer progressivement les détails afin de simuler d'éventuels changements d'échelles. La figure 2.12 en donne une représentation.

$\sigma = 0$ $\sigma = 3$ $\sigma = 6$

FIG. 2.12 – Représentation multi-échelle d'une image, les détails disparaissent progressivement, en fonction de l'échelle d'observation. L'espace d'échelle permet donc d'observer la scène de plus en plus loin.

D'un point de vue mathématique, cet espace est défini comme étant l'ensemble des convolutions du signal par un filtre passe bas. Koenderink préconise l'utilisation d'un filtre à noyau gaussien. L'espace d'échelles linéaire (ou gaussien), noté L, est donc déterminé par :

$$L(\mathbf{x};\sigma) = g_\sigma * I(\mathbf{x}) \tag{2.27}$$

Il énumère également les différentes invariances gérées, et notamment :

– l'invariance à la translation :

$$(T_{(\Delta_x,\Delta_y)} g_\sigma * I)(\mathbf{x}) = (g_\sigma * T_{(\Delta_x,\Delta_y)} I)(\mathbf{x}) \tag{2.28}$$

où $T_{(\Delta_x,\Delta_y)}$ est le vecteur translation.

– l'invariance à la rotation :

$$R_\theta g_\sigma(\mathbf{x}) = g_\sigma(x\cos\theta + y\sin\theta, -x\sin\theta + y\cos\theta), \quad \forall\theta \in \mathbb{R} \tag{2.29}$$

où θ est l'angle de rotation.

Les dérivées partielles sont définies de la façon suivante :

$$L_{x^m y^n}(x,y,\sigma) = (\frac{\partial^{m+n}}{\partial x^m \partial y^n} g_\sigma * I)(x,y), \quad \forall\sigma \in \mathbb{R}^+ \tag{2.30}$$

De plus, il est possible de définir l'espace d'échelles laplacien, noté $\nabla^2 L$, déterminé par l'équation :

$$\nabla^2 L = L_{x^2} + L_{y^2} \tag{2.31}$$

Par la suite, Lindeberg [67][68] définit les opérateurs différentiels par différences finies suivant :

$$L_i[i,j;\sigma] = \frac{1}{2}(L[i+1,j;\sigma] - L[i-1,j;\sigma]) \tag{2.32}$$

$$L_{i^2}[i,j;\sigma] = L[i+1,j;\sigma] - 2L[i,j;\sigma] + L[i-1,j;\sigma] \tag{2.33}$$

permettant notamment, d'obtenir une approximation discrète de l'espace d'échelles. Il établie également la même année une méthode de réprésentation multi-échelles de l'image. Il propose enfin, en 1998 [70], une méthode de sélection automatique de l'échelle caractéristique, basée sur la maximisation du laplacien normalisé. D'un point de vue mathématique, en se basant sur l'équation 2.31, il est possible de normaliser l'espace d'échelles :

$$\nabla^2_{norm} L(\mathbf{x};\sigma) = \sigma^2 \nabla^2 L(\mathbf{x};\sigma) \tag{2.34}$$

et ainsi de déterminer l'échelle caractéristique :

$$\sigma_c(\mathbf{x}) = \underset{\sigma}{\operatorname{argmax}}(\nabla^2_{norm} L(\mathbf{x};\sigma)). \tag{2.35}$$

La figure 2.13 présente un exemple de détermination d'échelle caractéristique. Les courbes caractérisent le laplacien normalisé de l'image initiale (à gauche) et de l'image ayant subi un changement d'échelle de $\sigma = 3$ (à droite). Nous pouvons observer que le rapport des échelles caractéristiques extraites permet d'obtenir la valeur de ce σ.

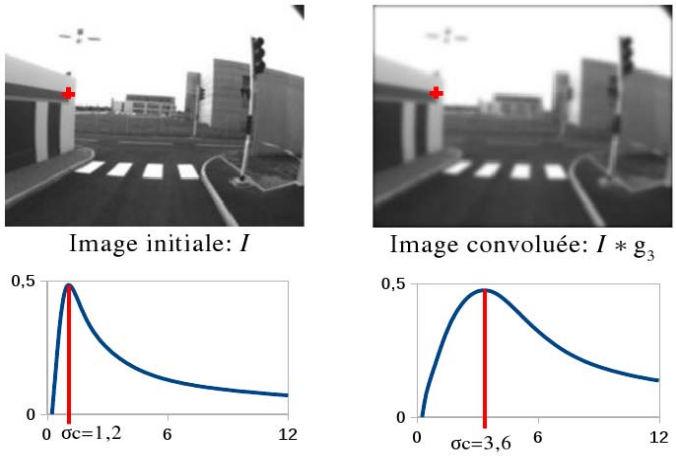


FIG. 2.13 – Représentation du laplacien normalisé en fonction de l'échelle d'exploration. Le maximum est atteint pour la valeur d'échelle caractéristique.

2.2.2.1 Détecteur Harris-Laplace

Afin d'accroître les performances et le nombre d'invariances du détecteur de Harris, Mikolajczyk et Schmid [80] proposent en 2001, d'y intégrer cette notion d'espace d'échelles et de sélection d'échelle caractéristique. Plusieurs méthodes ont alors été envisagées :

– le gradient carré :

$$\sigma^2(L_x^2(\mathbf{x};\sigma) + L_y^2(\mathbf{x};\sigma)) \tag{2.36}$$

– le laplacien (LoG) :

$$\sigma^2(L_{xx}(\mathbf{x};\sigma) + L_{yy}(\mathbf{x};\sigma)) \tag{2.37}$$

– la différence de gaussienne (DoG) :

$$I * (g_{\sigma_{n-1}}(\mathbf{x}) - g_{\sigma_n}(\mathbf{x})) \tag{2.38}$$

Le tableau suivant relève les résultats des tests effectués sur les méthodes énoncées :

	LoG	DoG	Gradient carré
Points détectés	46%	38%	30%
Taux bonnes détections	29%	28%	22%
Bonnes détections finales	13,3%	10,6%	6,6%

TAB. 2.1 – Tests des différents estimateurs d'espace d'échelle.

Au vu de ces résultats, Mikolajczyk et Schmid proposent une approche multi-échelle du détecteur de Harris, basée sur le Laplacien. L'équation 2.17 est modifiée afin d'obtenir l'équation suivante :

$$\mathbf{M} = \sigma_D^2 g_{\sigma_I} * \begin{bmatrix} L_x^2(\mathbf{x};\sigma_D) & L_x L_y(\mathbf{x};\sigma_D) \\ L_x L_y(\mathbf{x};\sigma_D) & L_y^2(\mathbf{x};\sigma_D) \end{bmatrix}, \tag{2.39}$$

où σ_D et σ_I représentent respectivement les échelles de différenciation et d'intégration. La relation liant ces deux valeurs est définie par :

$$\sigma_D = k \times \sigma_I \qquad avec \qquad k \in [0,5; 0,75] \tag{2.40}$$

Mikolajczyk et Schmid préconise une valeur égale à 0,7 pour le coefficient k. L'extraction des points d'intérêt se divise en deux parties :

– la détection de points d'intérêt, par le biais des équations 2.19 et 2.39, pour n valeurs d'échelle. Ceci permet d'obtenir les cartes de Harris dans l'espace multi-échelles (exemple de carte de Harris : figure 2.14).

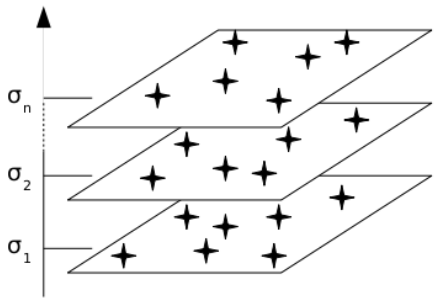


FIG. 2.14 – Création des cartes de Harris pour n échelles.

– la détermination de l'échelle caractéristique. En effet, il apparaît qu'à certains points de l'image initiale, correspondent plusieurs points à des échelles différentes. Il s'agit alors de déterminer ceux qui caractérisent réellement des structures locales. Il faut donc définir, pour chaque point détecté, son laplacien normalisé, afin de le maximiser pour en déduire l'échelle caractéristique (équation 2.35).

Les points d'intérêt ainsi obtenus sont donc définis par leurs coordonnées (x, y) et par leur échelle caractéristique (ou d'intégration) notée σ_I.

2.2.2.2 Détecteur basé sur des différences de gaussiennes (DoG)

Lowe [73] propose en 1999, d'intégrer la notion d'espace d'échelles dans le calcul de différences de gaussiennes, afin de rendre plus stable la détection de points d'intérêt. En se basant sur l'équation 2.27, il définit la fonction $D(x, y; \sigma)$ comme étant la différence de deux espaces d'échelles gaussiens consécutifs :

$$D(x, y; \sigma) = L(x, y; k\sigma) - L(x, y; \sigma) \tag{2.41}$$

avec k un facteur multiplicateur constant (généralement égal à $\sqrt{2}$). La figure 2.15 permet de visualiser les différentes étapes de construction du DoG.

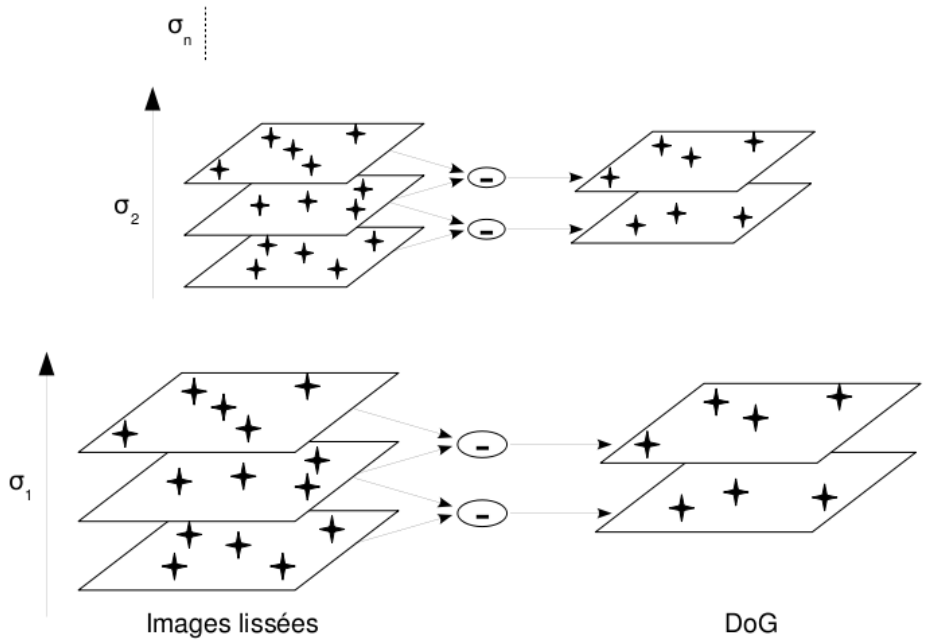


FIG. 2.15 – Schéma des différences de gaussiennes : Partie gauche : ensemble des images lissées par une gaussienne. Partie droite : ensemble des images des différences de gaussiennes.

L'image résultante de la différence de gaussiennes peut être illustrée par la figure 2.16.

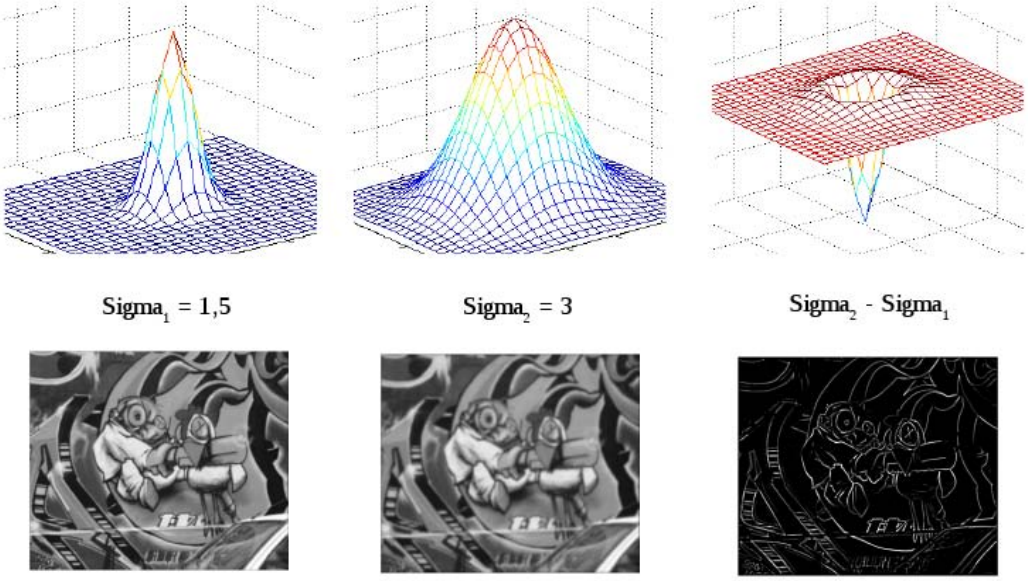


FIG. 2.16 – Illustration d'une différence de gaussiennes.

Cette approche gère la notion d'espace d'échelle, par le biais des n octaves, à l'intérieur desquels l'image est convoluée par une gaussienne d'écart type $k\sigma$. Les images résultantes des différences de gaussiennes permettent de déterminer les maxima locaux. En effet, un point est dit d'intérêt si sa valeur est maximale sur sa 26-connexité (figure 2.17).

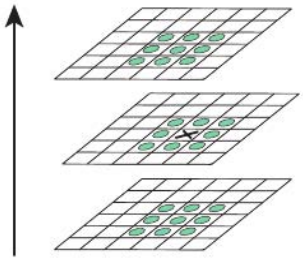


FIG. 2.17 – DoG : sélection des points d'intérêt suivant sa 26-connexité (image extraite de [73]).

En 2004, Lowe [74] apporte quelques précisions sur cette méthode. Il démontre notamment la relation entre la fonction D (équation 2.41) et le calcul du laplacien normalisé (équation 2.34). En partant de l'équation de diffusion :

$$\frac{\partial L}{\partial \sigma} = \sqrt{t}\nabla^2 L \tag{2.42}$$

où $t = \sigma^2$. Il est possible d'approximer cette dérivée partielle :

$$\frac{\partial L}{\partial \sigma} \approx \frac{L(x, y, k\sigma) - L(x, y, \sigma)}{k\sigma - \sigma} \tag{2.43}$$

et d'en déduire ainsi la relation suivante :

$$L(x, y, k\sigma) - L(x, y, \sigma) \approx (k-1)\sigma^2 \nabla^2 L \tag{2.44}$$

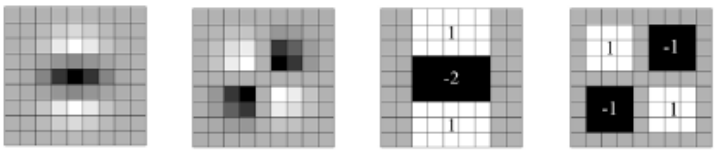
L'espace d'échelles normalisé ainsi obtenu correspond à celui énoncé par Lindeberg, à un facteur $(k-1)$ près. Lowe redémontre donc que la sélection d'échelle caractéristique est faite par le biais des n octaves. Il précise également que le facteur $(k-1)$ n'a aucun impact sur la stabilité de la détection des points d'intérêt.

2.2.2.3 Détecteur fast-hessien

En 2006, Bay et al. [13] proposent un détecteur, basé sur une approximation du filtrage gaussien, leur permettant de diminuer considérablement les temps de calculs. En se basant sur l'équation 2.1, on obtient pour un point $\mathbf{x}$ à une échelle σ :

$$\mathbf{H}_\sigma(\mathbf{x}) = \begin{bmatrix} L_{xx}(\mathbf{x};\sigma) & L_{xy}(\mathbf{x};\sigma) \\ L_{xy}(\mathbf{x};\sigma) & L_{yy}(\mathbf{x};\sigma) \end{bmatrix}. \tag{2.45}$$

Bay et al. approximent les dérivées secondes des gaussiennes par des filtres plus simples présentés en figure 2.18.



a b c d

FIG. 2.18 – Représentation des dérivées secondes des fonctions gaussiennes (a, b) et des filtres d'approximations (c, d) utilisés par la méthode SURF.

Les convolutions otenues par des dérivées régularisées (par un filtrage gaussien) sont notées D_{xx}, D_{xy} et D_{yy}. Afin de garder la cohérence dans le filtrage, Bay et al. utilisent un filtre approximé initial de taille 9×9 correspondant à un filtre gaussien d'écart type $\sigma = 1,2$. Cette cohérence est assurée par l'équation suivante :

$$\frac{|L_{xy}(1,2)|_F |D_{xx}|_F}{|L_{xx}(1,2)|_F |D_{xy}|_F} = 0,912... \approx 0,9 \tag{2.46}$$

où $|\bullet|_F$ est la norme de Frobenius. Le hessien sera donc déterminé par l'équation suivante :

$$\det(\mathbf{H}_{approx}) = D_{xx}D_{yy} - (0,9D_{xy})^2 \tag{2.47}$$

Enfin, afin de gérer le multi-échelle, Bay et al. s'appuient sur un ensemble de masques de tailles croissantes (9×9, 15×15, 21×21, 27×27,...) dépendant de l'écart type de la gaussienne à approximer. Par exemple dans le cas d'un filtre de taille 27×27, l'approximation correspond à une gaussienne possédant un $\sigma = \frac{27}{9} \times 1,2 = 3,6$.

Ce détecteur permet notamment la détection de zones homogènes comme le montre la figure 2.19.

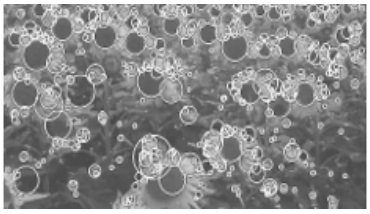


FIG. 2.19 – Détection de zones homogènes par le fast-hessien (image extraite de [13]).

2.2.3 Invariances aux transformations affines et projectives

Afin d'être le plus robuste possible aux transformations affines, des adaptations des différentes méthodes étudiées précédemment ont été proposées. Introduit en 1994 par Lindeberg [69] puis détaillé en 1997 par Lindeberg et Garding [71], cette adaptation se base sur l'interprétation de la matrice des moments seconds notée μ_L (ou matrice d'auto-correlation) en multi-échelle, définie par :

$$\mu_L(\mathbf{x}; \Sigma_t, \Sigma_s) = g_{\Sigma_s} * (\nabla_L(\mathbf{x}; \Sigma t) \nabla_L^T(\mathbf{x}; \Sigma t)), \tag{2.48}$$

où $\nabla_L = (L_x, L_y)^T$, Σ_t et Σ_s correspondent aux matrices de covariance. Lindeberg et Garding démontrent également que pour toute transformation affine $\mathbf{x}_1 = \mathbf{B}\mathbf{x}_2$, où $\mathbf{B}$ représente la transformation affine, l'adaptation de la matrice des moments seconds en multi-échelle est donnée par :

$$\mu_L(\mathbf{x}_1; \Sigma_t, \Sigma_s) = \mathbf{B}^T \mu_R(\mathbf{x}_2; \mathbf{B}\Sigma_t\mathbf{B}^T, \mathbf{B}\Sigma_s\mathbf{B}^T)\mathbf{B}, \tag{2.49}$$

où μ_L et μ_R sont les matrices des moments seconds, centrées respectivement en $\mathbf{x}_1$ et $\mathbf{x}_2$. $\mathbf{B}$ peut donc être estimée en interprétant l'égalité 2.49. La conséquence, énoncée par Lindeberg, est que s'il existe une transformation affine $\mathbf{B}$ de telle sorte que μ_R reste constante et égale à la matrice identité, alors le point observé est invariant aux transformations affines.

D'un point de vue implémentation, cette méthode se détaille en quatre étapes :
- estimation locale de la matrice des moments seconds μ ;
- estimation locale de la transformation affine $\mathbf{B}$, proportionnelle à $\sqrt{\mu}$;
- transformation de l'image initiale par $\mathbf{B}$;
- itération jusqu'à l'obtention d'une matrice μ constante.

La figure 2.20 propose une illustration de cette méthode sur une région circulaire. Nous observons qu'à partir de la quatrième itération, l'estimation de la matrice μ devient stable et constante.

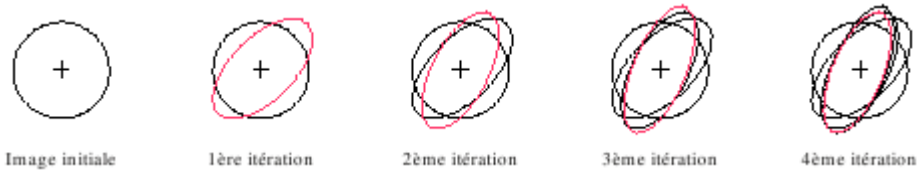


FIG. 2.20 – Exemple de détermination d'une transformation affine, de façon itérative.

Baumberg propose [12] en 2000, un procédé couplant une méthode de détection, une adaptation affine et une mise en correspondance. Cette méthode sera reprise et améliorée, quatre ans plus tard, afin de créer les détecteurs Harris-affine de Hessian-affine.

2.2.3.1 Détecteurs Harris-affine et Hessian-affine

En 2004, Mikolajczyk et Schmid proposent de coupler un détecteur multi-échelles avec la méthode d'adaptation (ou de normalisation) affine décrite précédemment. Ils proposent ainsi deux nouveaux détecteurs [82] : le Harris-affine et le Hessian-affine. Pour chacun d'entre eux, la méthode repose tout d'abord sur l'extraction multi-échelles des points d'intérêt, puis sur la détermination itérative d'une région locale circulaire.

La partie analyse multi-échelles de la scène est assurée par l'utilisation du détecteur Harris-Laplace (équation 2.39) pour le Harris-affine, et celle du détecteur Hessian-Laplace (équation 2.45) pour le Hessian-affine.

La détermination de la région elliptique centrée sur le point d'intérêt se divise en plusieurs étapes. Dans un premier temps l'initialisation de la matrice de transformation $\mathbf{U}^0$ (correspondant à l'identité) et la récupération des données extraites ($\mathbf{x}^0$, σ_D^0 et σ_I^0) est nécessaire. L'application de la matrice $\mathbf{U}^k$ sur l'image permettra de la déformer afin de faire converger la région locale vers une forme circulaire. L'étape suivante consiste à déterminer les nouvelles valeurs d'échelle d'intégration σ_I^k et de différentiation σ_D^k en se basant sur les équations suivantes :

$$\sigma_I^k = \underset{\sigma_I = t\sigma_I^{(k-1)} \text{ pour } t\in[0,7;1,4]}{\operatorname{argmax}} (\sigma_I^2 \times det(L_{xx}(\mathbf{x};\sigma_I) + L_{yy}(\mathbf{x};\sigma_I))) \tag{2.50}$$

$$\text{et} \quad \sigma_D^k = \underset{\sigma_D = s\sigma_I^k \text{ pour } s\in[0,5;0,75]}{\operatorname{argmax}} \frac{\lambda_{min}(\mu(\mathbf{x}^k;\sigma_I^k,\sigma_D))}{\lambda_{max}(\mu(\mathbf{x}^k;\sigma_I^k,\sigma_D))} \tag{2.51}$$

En utilisant ces deux valeurs dans l'équation 2.19, Mikolajczyk et Schmid proposent un recalage du point d'intérêt $\mathbf{x}^0 \rightarrow \mathbf{x}^k$ et le définissent par :

$$\mathbf{x}^k = \underset{\mathbf{x}\in W(\mathbf{x}^{k-1})}{\operatorname{argmax}} (\det(\mu(\mathbf{x};\sigma_I^k,\sigma_D^k)) - \alpha \text{trace}^2(\mu(\mathbf{x};\sigma_I^k,\sigma_D^k))) \tag{2.52}$$

Les étapes suivantes consistent à mettre à jour la matrice $\mathbf{U}^k = \mu^k \mathbf{U}^{k-1}$ et à réitérer ce processus jusqu'à obtention d'une région locale circulaire. La figure 2.21 illustre cette méthode d'adaptation affine.

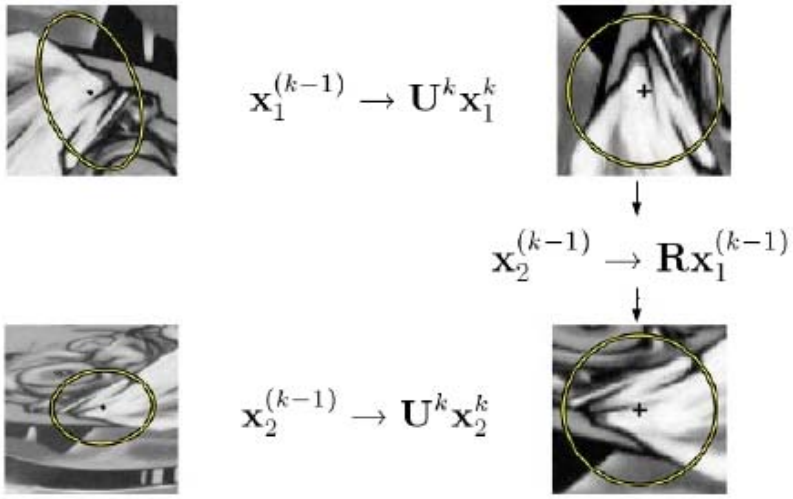


FIG. 2.21 – Exemple d'adaptation affine d'une région elliptique. Le calcul itératif de la matrice $\mathbf{U}$ permet de ramener le problème à une simple rotation.

La normalisation affine, caractérisée par la matrice $\mathbf{U}^k$ entraine la transformation de l'ellipse en cercle. Les deux régions obtenues sont liées par une simple rotation $\mathbf{R}$, permettant ainsi la suppression des problèmes dus aux transformations affines et projectives. En définitive, ces deux méthodes ne diffèrent que par leur partie analyse multi-échelles, et donnent un nombre de points et des résultats similaires.

2.2.3.2 Détecteur MSER

Matas et al. [78] proposent en 2002 une approche originale permettant de définir une région d'intérêt robuste aux transformations affines. Dans un premier temps, une étape de classification permet de regrouper en classe chaque pixel de l'image suivant sa valeur d'intensité. L'histogramme ainsi créé permet de déterminer une fonction d'intensité en se basant sur l'aire de chaque classe. En appliquant différents seuillages à cette fonction d'intensité, certaines régions de l'image classifiée vont varier et d'autres non. Les régions d'intérêt correspondent donc à celles qui restent robustes aux différents seuillages. Matas et al. démontrent que les régions d'intérêt ainsi extraites, sont invariantes aux transformations affines, aussi bien photométriques que géométriques.

2.3 Les différentes méthodes de description d'un point d'intérêt

Afin de permettre la mise en relation des différents points détectés, l'utilisation d'une méthode de description est indispensable. Cette dernière permet de caractériser chaque point d'intérêt, et d'en extraire différentes composantes (intensités, informations sur le voisinage, échelles, gradients, ...). N'ayant proposé que des détecteurs locaux dans le paragraphe précédent, l'étude des descripteurs se limitera donc au domaine local.

Il existe une multitude de méthodes de description, chacune ayant ses avantages et ses inconvénients. Il est donc indispensable d'en connaître les caractéristiques afin de choisir la plus appropriée à la problématique. D'un point de vue historique, la première méthode à avoir été proposée se base uniquement sur le point d'intérêt, et plus particulièrement sur l'utilisation de ses coordonnées. Par la suite l'observation de son voisinage a permis d'améliorer fortement les résultats. Pour ce faire, un certain nombre de méthodes ont été créées se basant sur un résumé visuel plutôt que sur les pixels, nous pouvons citer les moments, les transformées, ou encore les histogrammes. Nous proposons donc de détailler ces différents descripteurs et leur méthode de caractérisation.

2.3.1 Descripteur basé sur les moments

La description basée sur les moments a été initialement proposée pour de la reconnaissance d'objets. Ils étaient utilisés pour travailler sur des images contenant l'objet entier. L'avantage des moments, tels que les moments de Hu et de Zernike, est leur invariance aux translations, rotations et changements d'échelle isotrope. Néanmoins, un changement de point de vue, un changement d'échelle anisotrope ou l'ajout d'occultations, provoque une forte diminution de la qualité des résultats. En se basant sur les études comparatives proposées par Choksuriwong et al. en 2005 [26] et 2008 [27], nous proposons d'en détailler les principaux composants.

2.3.1.1 Les moments de Hu

Introduits en 1962 par Hu [53], leur utilisation était tout d'abord globale, puis en limitant le calcul des moments à un voisinage d'un point d'intérêt, cette méthode est devenue locale. Cette dernière se base sur la détermination des moments centraux et de leur normalisation. Soit un moment d'ordre $(p+q)$ avec $(p,q>0)$, l'équation 2.53 détaille la détermination du moment central.

$$m_{p,q}(x_0,y_0)=\int_{\Omega} x^p y^q I(x+x_0;y+y_0)dxdy, \tag{2.53}$$

où (xo,yo) représentent les coordonnées du point d'intérêt. L'équation 2.54 caractérise la normalisation de ce moment.

$$\mu_{p,q}=\frac{m_{p,q}}{m_{0,0}^{(1+(p+q)/2)}} \tag{2.54}$$

Hu propose donc d'utiliser une succession de polynômes, basés sur l'équation 2.54, afin de créer un descripteur local. Les polynômes présentés en 2.55 caractérisent ainsi le voisinage du point d'intérêt.

$$\begin{cases} M_{Hu1} = \mu_{2,0} + \mu_{0,2} \\ M_{Hu2} = (\mu_{2,0} - \mu_{0,2})^2 + 4\mu_{1,1}^2 \\ M_{Hu3} = (\mu_{3,0} - 3\mu_{1,2})^2 + (3\mu_{2,1} - \mu_{0,3})^2 \\ M_{Hu4} = (\mu_{3,0} + \mu_{1,2})^2 + (\mu_{2,1} + \mu_{0,3})^2 \\ M_{Hu5} = (\mu_{3,0} - 3\mu_{1,2})(\mu_{3,0} + \mu_{1,2})[(\mu_{3,0} + \mu_{1,2})^2 - 3(\mu_{2,1} + \mu_{0,3})^2] \\ \qquad + (3\mu_{2,1} + \mu_{0,3})(\mu_{2,1} + \mu_{0,3})[3(\mu_{3,0} + \mu_{1,2})^2 - (\mu_{2,1} + \mu_{0,3})^2] \\ M_{Hu6} = (\mu_{2,0} - \mu_{0,2})[(\mu_{3,0} + \mu_{1,2})^2 - (\mu_{2,1} + \mu_{0,3})^2] + 4\mu_{1,1}(\mu_{3,0} + \mu_{1,2}) \\ \qquad (\mu_{2,1} + \mu_{0,3}) \\ M_{Hu7} = (3\mu_{2,1} - \mu_{0,3})(\mu_{3,0} + \mu_{1,2})[(\mu_{3,0} + \mu_{1,2})^2 - 3(\mu_{2,1} + \mu_{0,3})^2] \\ \qquad - (\mu_{3,0} - 3\mu_{1,2})(\mu_{2,1} + \mu_{0,3})[3(\mu_{3,0} + \mu_{1,2})^2 - (\mu_{2,1} + \mu_{0,3})^2] \end{cases}, \quad (2.55)$$

M_{Huk} (avec $k \in [\![1; 7]\!]$) représente le moment de Hu d'indice k. L'auteur décrit et justifie également les invariances par translation, rotation et changement d'échelle de son descripteur.

2.3.1.2 Les moments de Zernike

Les polynômes de Zernike ont été introduits en 1934 avant d'être utilisés dans le domaine de l'optique, de la robotique, puis en vision par ordinateur, notamment en 2003 par Chong et al. [28]. D'un point de vue général, un polynôme de Zernike, noté $P(r, \theta)$ et caractérisé par son rayon r et son angle θ, est défini par :

$$P_{mn}(r, \theta) = R_{mn}(r)e^{-jn\theta}, \quad (2.56)$$

où m et n représentent l'ordre du moment, et $R_{mn}(r)$ le polynôme radial orthogonal défini par :

$$R_{mn}(r) = \sum_{s=0}^{\frac{m-|n|}{2}} (-1)^s \frac{(m-s)!}{s!(\frac{m+|n|}{2} - s)!(\frac{m-|n|}{2} - s)!} r^{m-2s} \quad (2.57)$$

Afin de décrire localement un point d'intérêt, l'auteur propose d'utiliser les moments de Zernike. Ces derniers se basent sur un ensemble de polynômes décrits par l'équation 2.56, caractérisant ainsi un ensemble orthogonal défini sur un disque unité. Il est par conséquent possible de déterminer les moments de Zernike de la façon suivante :

$$M_{Zmn} = \frac{m+1}{\pi} \int_{\Omega^2} \mathrm{V}_I(\mathbf{x}; \Omega)[P_{mn}(r, \theta)] dr d\theta. \quad (2.58)$$

Finalement, le descripteur est composé de seize moments de Zernike, suivant les seize polynômes listés en 2.59.

$$\begin{cases} P_{00}(r,\theta) = 1 & P_{11}(r,\theta) = r \\ P_{02}(r,\theta) = 2r^2 - 1 & P_{22}(r,\theta) = r^2 \\ P_{13}(r,\theta) = 3r^3 - 2r & P_{33}(r,\theta) = r^3 \\ P_{04}(r,\theta) = 6r^4 - 6r^2 + 1 & P_{24}(r,\theta) = 4r^4 - 3r^2 \\ P_{44}(r,\theta) = r^4 & P_{15}(r,\theta) = 10r^5 - 12r^3 + 3r \\ P_{35}(r,\theta) = 5r^5 - 4r^3 & P_{55}(r,\theta) = r^5 \\ P_{06}(r,\theta) = 20r^6 - 30r^4 + 12r^2 - 1 & P_{26}(r,\theta) = 15r^6 - 20r^4 + 6r^2 \\ P_{46}(r,\theta) = 6r^6 - 5r^4 & P_{66}(r,\theta) = r^6 \end{cases} \tag{2.59}$$

Les résultats proposés dans la littérature montrent que l'utilisation de ce type de moment est meilleure, notamment en termes de redondance de l'information et de possibilité de reconstruction.

2.3.2 Descripteur basé sur les transformées intégrales

Il existe un certain nombre de méthodes utilisant les transformées intégrales pour décrire localement un point d'intérêt. Ghorbel en donne une description succincte en 1994 [44], puis Derrode et al. s'intéressent à l'utilisation des transformées de Fourier en 1999 [35] et 2001 [36]. Plus récemment Mennesson et al. [79] présentent de nouveaux descripteurs utilisant notamment une généralisation des transformées. Le principe de ces méthodes repose sur l'interprétation de la transformée notée T définie par :

$$I(\mathbf{x}) \overset{\mathrm{T}}{\leftrightarrow} \hat{I}(\nu) = \int_{\Gamma} I(\mathbf{x}) k(\mathbf{x};\nu) d\mathbf{x} \tag{2.60}$$

où k représente le noyau de la transformée. Cette dernière doit remplir deux conditions, elle doit être inversible et le module de la transformée doit être invariant.

En se basant sur la relation 2.60, la transformée de Fourier locale d'un point d'intérêt est égale à :

$$T_F(\nu_x,\nu_y) = \int_{\mathbb{R}^2} \mathrm{V}_I(\mathbf{x};\Omega) e^{-2i\pi(x\nu_x + y\nu_y)} dxdy, \tag{2.61}$$

Il est également possible d'utiliser la transformée de Fourier circulaire, définie par l'équation suivante :

$$T_{Fc}(r,\nu_\theta) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{V}_I(r,\theta;\Omega) e^{-2i\pi\theta\nu_\theta} d\theta, \tag{2.62}$$

où r et θ correspondent aux coordonnées polaires du point.

Derrode et al. proposent également l'utilisation de la transformée de Fourier-Mellin définie par :

$$T_{FM}(\nu_x,\nu_y) = \int_0^{+\infty} \int_0^{+\infty} \mathrm{V}_I(\mathbf{x};\Omega) x^{i(\nu_x-1)} y^{i(\nu_y-1)} dxdy \tag{2.63}$$

En couplant ces différents types de transformées, les auteurs obtiennent des invariances par translation (avec Fourier), par rotation (avec Fourier circulaire et Fourier-Mellin) et par changement d'échelles (avec Fourier-Mellin).

2.3.3 Descripteur basé sur les histogrammes

D'un point de vue général, l'histogramme représente une estimation de la distribution des intensités de l'image. Swain et Ballard [106] utilisent en 1991 des histogrammes ayant pour objectif la reconnaissance d'objets. De nombreuses méthodes ont dès lors été proposées et il est posible de les classer en deux catégories, l'une s'appuyant sur des histogrammes d'intensités lumineuses [106] [98] et l'autre sur des histogrammes de gradients orientés [22] [74] [73] [46] [13] [108] [86] [32] [107]. Nous proposons d'en donner une description afin de faciliter le choix de la méthode à utiliser.

2.3.3.1 Histogramme d'intensité lumineuse (ou de couleur)

L'histogramme de couleur proposé en 1991 par Swain et Ballard [106] puis repris par Schiele et Waibel [98] en 1995, est utilisé comme résumé visuel de l'image. Il présente l'avantage d'avoir une construction rapide et peu onéreuse en terme d'espace mémoire. Appelé communément 'estimateur de densité non-paramétrique du premier ordre', l'histogramme consiste en un graphique statistique permettant de représenter la distribution des intensités des pixels. Généralement appliqué à l'image entière, il est néanmoins possible de l'extraire dans le voisinage d'un point d'intérêt. L'histogramme h du voisinage d'un point est défini par :

$$h_{\mathbf{x}}(i) = \frac{1}{\text{card}(\Omega)} \sum_{\mathbf{x} \in \Omega} \mathbb{1}(I(\mathbf{x}) = i) \qquad \text{avec } i \in Q = [\![0; 255]\!]. \tag{2.64}$$

Pour résumer, l'histogramme consiste donc à comptabiliser le nombre de pixels présentant la même valeur d'intensité dans le voisinage considéré.

2.3.3.2 Histogramme de gradients orientés (HOG)

Proposés en 2005 par Dalal et Triggs [32], les histogrammes de gradients orientés sont utilisés principalement en vision par ordinateur pour de la détection d'objets. Leurs utilisations se sont également révélées particulièrement efficaces pour la détection de personnes. D'un point de vue général, un gradient permet de calculer les variations d'une fonction par rapport aux changements de ses paramètres. Dans le cas d'une image, la détermination du gradient consiste à calculer la variation de l'intensité des pixels dans différentes directions. La figure 2.22 illustre le calcul des gradients horizontaux et verticaux sur une image entière.

FIG. 2.22 – Exemple de gradients d'une image. A gauche : image initiale, au centre : gradients horizontaux, à droite : gradients verticaux.

En effectuant une telle analyse sur une image entière ou sur un voisinage, nous construisons un histogramme de gradients orientés permettant d'étudier les orientations des gradients locaux. L'idée directrice d'un HOG est que l'apparence et la forme d'un objet dans une image peuvent être décrites par la répartition de l'intensité du gradient. L'analyse des différentes méthodes existantes permet de différencier deux types d'architecture : une construction à base carrée appelée R-HOG et une autre à base circulaire appelée C-HOG. Les différents descripteurs utilisant des histogrammes de gradients orientés peuvent donc être classés dans l'une des deux catégories citées.

A) R-HOG

a) SIFT

Nous avons présenté au §2.2.2.2 le détecteur proposé par Lowe, utilisant des différences de gaussiennes afin d'extraire les primitives. L'auteur propose également en 1999 [73] puis en 2004 [74], de coupler ce détecteur à un descripteur local basé sur des histogrammes de gradients orientés. Les données fournies initialement par le détecteur sont les coordonnées des points d'intérêt ainsi que leur échelle caractéristique. Une étape préliminaire consiste à construire l'histogramme des orientations locales définies en chaque point ($\mathbf{x}$) par :

$$\theta(\mathbf{x}) = \arctan\left(\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)}\right) \tag{2.65}$$

Cet histogramme se compose de trente-six intervalles (36 classes), couvrant chacun un angle de dix degrés. Ce dernier est pondéré d'une part par un filtre gaussien d'écart-type égal à une fois et demie la valeur de l'échelle locale et d'autre part par l'amplitude m de chaque point defini par :

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \tag{2.66}$$

Les pics ainsi obtenus correspondent aux orientations dominantes. L'auteur préconise un seuillage de la sélection de ces orientations afin qu'elles permettent d'atteindre au moins quatre-vingts pour cent de la valeur maximale. La figure 2.23 représente la construction d'un tel histogramme et la sélection des orientations dominantes.

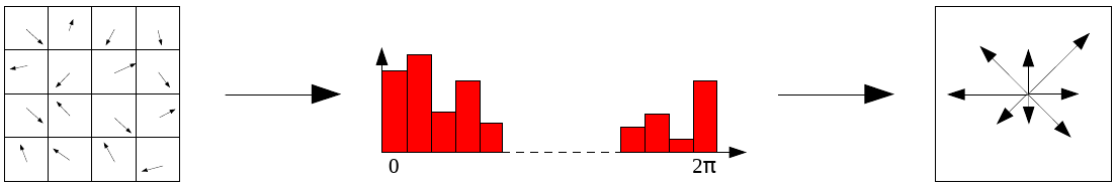


FIG. 2.23 – Exemple d'histogramme des orientations. A gauche : gradients des pixels définissant le voisinage du point, au centre : histogrammes des orientations suivant 36 bins, à droite : extraction des orientations dominantes.

Par souci de visibilité nous ne représentons que huit des trente-six orientations possibles sur la partie droite de la figure 2.23. En définitive, chaque point d'intérêt est défini par quatre paramètres : ses coordonnées, son échelle et son orientation.

La fenêtre de description du voisinage du point d'intérêt possède une taille fixe de 16x16 pixels, subdivisée en 4x4 zones de 4x4 pixels chacune. Dans un premier temps ce masque d'analyse est recalé, par le biais d'une rotation d'un angle égal à l'orientation local du point, afin de garantir l'invariance à la rotation. A l'intérieur de chacune des seize zones est calculé un histogramme des orientations basé sur huit intervalles (suivant un angle de $\frac{\pi}{4}$) et subissant deux pondérations, une par l'amplitude du gradient et l'autre par convolution avec une gaussienne. Le choix de huit classes pour la répartition des gradients s'appuie sur un certain nombre de tests et de résultats. Nous proposons de visualiser l'influence du nombre d'intervalles sur les performances du descripteur (figure 2.24).

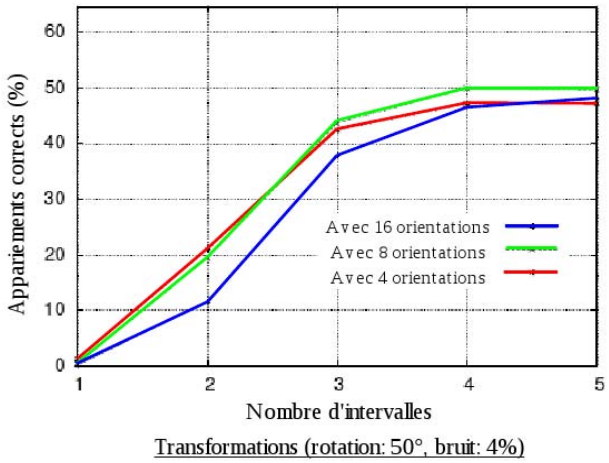


FIG. 2.24 – Graphique extrait de [73] représentant l'influence du nombre d'intervalles sur les performances du descripteur. Au vue de ces courbes, le choix de huit classes présente les meilleurs résultats.

L'étape suivante consiste à concaténer et normaliser les seize histogrammes ainsi obtenus. Afin de limiter la sensibilité du descripteur aux changements de luminosité, les valeurs inférieurs à 0,2 sont remplacées par 0 et l'histogramme est de nouveau normalisé. La figure 2.25 illustre la construction de ce descripteur.

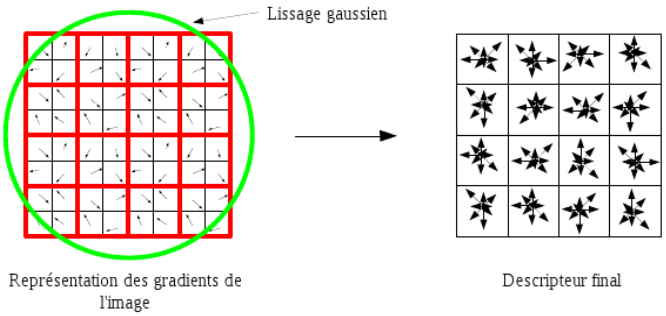


FIG. 2.25 – (Gauche) : le lissage et l'analyse des gradients locaux, (droite) : la concaténation des histogrammes de zones afin de créer le descripteur final.

La dimension de ce descripteur a tout d'abord été perçue comme étant trop grande, néanmoins Lowe démontre que le choix d'un descripteur de taille 128 (seize histogrammes de huit classes chacun) permet d'accroître fortement les performances tout en accusant une augmentation assez faible du temps de calcul. Il a également démontré que le taux d'appariement dépasse les 50% pour un changement de point de vue supérieur à 50°, lui permettant ainsi d'être robuste aux transformations affines d'un point de vue local.

En 2006, Grabner et al. [46] proposent une méthode d'approximation du SIFT afin d'en diminuer les temps de calculs. Se basant sur les travaux publiés en 2001 par Viola et Jones [109], les auteurs opèrent un certain nombre de modifications. Tout d'abord la détection s'effectue sur des images intégrales et n'interprète plus les différences de gaussiennes mais les différences des moyennes locales (DoM : Difference of Mean). La figure 2.26 compare la méthode SIFT 'classique' et celle approximée.

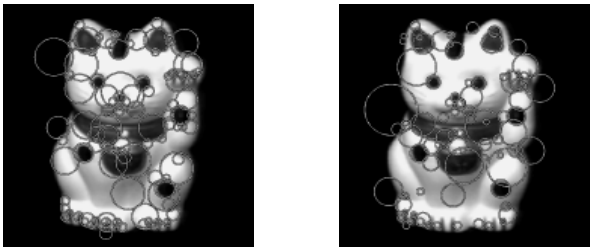


FIG. 2.26 – Comparaison entre une détection basée DoM (gauche) et une détection basée DoG (droite). (images extraites de [46])

Le taux de répétabilité entre les deux approches est sensiblement le même, tout comme le nombre de points détectés, mais le temps d'execution a été divisé par quatre. Afin d'améliorer également la partie description locale, Grabner et al. s'appuient sur l'utilisation d'histogrammes intégrals proposés par Porikli en 2005 [91]. Ce procédé permet de supprimer la dépendance entre la taille du masque d'analyse et la construction des histogrammes. En définitive, ces différentes approximations (ou modifications), permettent de diminuer les temps de calculs d'un facteur huit.

La même année, Chiu et Lozano-Perez [24] proposent de remplacer le masque descriptif carré du SIFT par une analyse circulaire adaptative. Le principe consiste donc à créer une zone circulaire autour du point d'intérêt, et de déformer ce cercle en se basant sur les informations locales fournies par le détecteur. Ce procédé permet d'accroître les performances du descripteur, notamment pour des problèmes de changements de point de vue.

b) SURF

En 2006, Bay et al. [13] proposent une nouvelle méthode de description locale de points d'intérêt, nommée SURF (Speeded-Up Robust Features). Fortement influencés par l'approche de SIFT, ils couplent une étape de recalage de la zone d'analyse avec la construction d'un histogramme de gradients orientés. La première étape de leur processus est donc de déterminer l'angle de rotation (ou de recalage) à appliquer à la fenêtre de description locale. Pour se faire, les auteurs appliquent des ondelettes de Haar sur l'image intégrale permettant ainsi de diminuer les temps de calculs de façon significative. Ces ondelettes permettent de calculer les dérivées premières de l'image sur un voisinage carré et d'étudier ainsi la répartition des gradients horizontaux et verticaux. Dès lors les réponses des ondelettes permettent de tracer le graphique de distribution des gradients et d'en déduire l'angle de recalage. La figure 2.27 schématise cette étape : sur l'image initiale le cercle représente la région d'intérêt dont le rayon est égal à $6s$ où s correspond à l'échelle caractéristique extraite du détecteur fast-hessien décrit au §2.2.2.3.

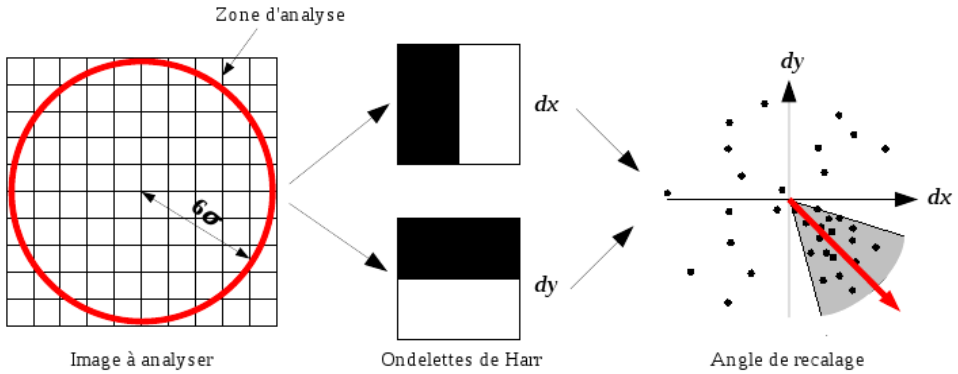


FIG. 2.27 – Détermination de l'angle de recalage du SURF, en analysant la répartition des réponses des ondelettes de Haar.

Les ondelettes de Haar sont constituées d'une partie noire ayant la valeur -1 et d'une partie blanche ayant la valeur +1 et leur taille est égale à $4s$. La détermination de l'angle de recalage illustrée par la figure 2.27 (partie de droite) se base sur la recherche de la répartition majoritaire des réponses des ondelettes dans une zone de rayon $\frac{\pi}{3}$ (zone grise sur le schéma).

La partie description locale se base quant à elle sur les sommes des réponses des ondelettes horizontales et verticales ainsi que sur leurs normes. Les figures 2.28 et 2.29 schématisent la construction du descripteur. Dans la première, nous représentons le masque d'analyse du SURF de taille $20s$ centrée sur le point d'intérêt.

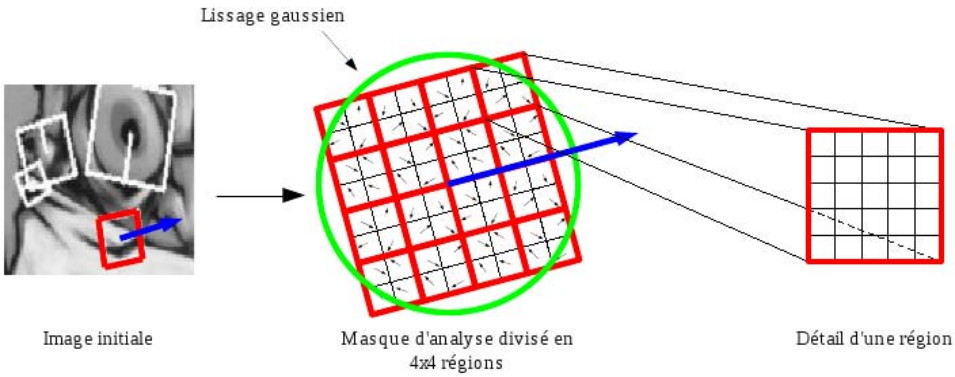


FIG. 2.28 – Masque d'analyse du SURF divisé en 4x4 régions, elles même divisées en 5x5 sous-régions. Le lissage gaussien est proportionnel à l'échelle locale : $\sigma = 3,3\sigma_c$.

Nous observons que la zone de description est divisée en seize régions, chacunes étant à leur tour échantillonées en vingt-cinq sous-régions. Une analyse en ondelettes est alors effectuée sur chaque région afin de construire le descripteur final, comme le montre le schéma de la figure 2.29.

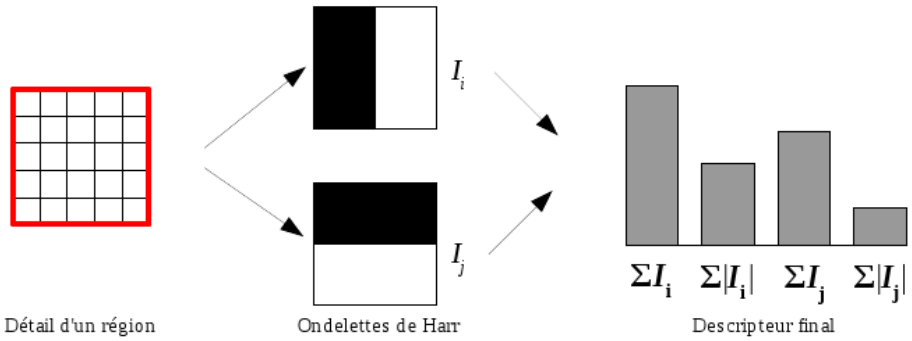


FIG. 2.29 – Extraction des différentes composantes du descripteur SURF par le biais des ondelettes de Haar dont la taille est égale à $2s$.

Le descripteur final est donc constitué de la somme des gradients en x et en y ainsi que de la somme de leur norme respective, pour l'ensemble des seize régions.

Bay et al. a proposé une variante à cette méthode (U-SURF) permettant de diminuer les temps de calculs mais entrainant une diminution de l'invariance à la rotation. Dans la même optique, Lepetit [19] a présenté en 2010 un descripteur binaire obtenant des résultats assez proches de ceux du SURF, tout en étant plus rapide.

c) CDIKP / KPB-SIFT

Présenté en 2008 par Tsai et al. [108] puis repris en 2010 par Zhao et al. [114], l'idée directrice est de remplacer l'étape d'analyse en ondelettes de Haar du SURF par l'utilisation de noyaux de Walsh-Hadamard (WH) [50]. Un certain nombre d'étapes, telles que l'extraction des points d'intérêt ou la sélection du voisinage, reste identique à celui des méthodes classiques (SIFT, SURF). Néanmoins la création de voisinages multi-résolution est une nouveauté. Elle permet d'extraire un voisinage à différentes échelles, ce procédé sera également utilisé par le descripteur de Cheng. Le recalage du masque d'analyse et le calcul des composantes du descripteur ont été modifiés et s'appuient sur les noyaux WH notés H_n et définis par :

$$H_n = \frac{1}{\sqrt{2}} \begin{bmatrix} H_{n-1} & H_{n-1} \\ H_{n-1} & -H_{n-1} \end{bmatrix} \qquad avec\ H_0 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{2.67}$$

Cet outil permet de créer un plus grand nombre de filtre et les tests démontrent l'accroissement du pouvoir discréminant du descripteur. Nous proposons d'illustrer (figure 2.30) la construction d'un tel descripteur.

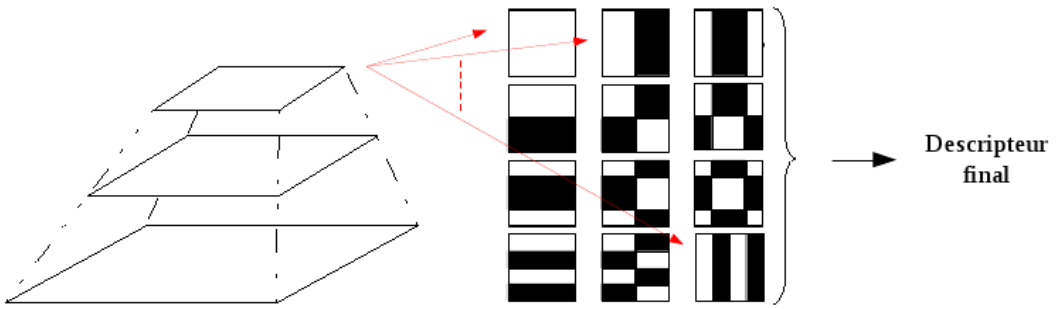


FIG. 2.30 – Exemple de noyaux de Walsh-Hadamard, où les parties noires sont égales à -1 et les parties blanches à 1. La partie de gauche représente les voisinages multi-résolutions du voisinage du point d'intérêt.

Chaque voisinage est filtré par l'ensemble des noyaux WH afin d'en extraire les différentes composantes constituant le descripteur final. Les auteurs montrent que leur descripteur possède une taille plus petite que celle du SIFT et permet d'accroître les performances de ce dernier.

d) A-SIFT

Introduit en 2009 par Morel et Yu [86][87], le ASIFT s'appuie sur les différents outils détaillés dans la méthode SIFT. L'objectif de leur approche est d'intégrer un modèle de caméra affine dans la méthode de description afin d'augmenter le nombre d'invariances ou de renforcer celles qui existent déjà. Les auteurs proposent donc d'estimer l'ensemble des transformations que peut subir l'image. Pour ce faire, ils utilisent un modèle de déformation défini pour un point $\mathbf{x}$ par :

$$\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{B} \qquad avec\ \mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\ et\ \mathbf{B} = \begin{bmatrix} e \\ f \end{bmatrix} \tag{2.68}$$

En identifiant la matrice $\mathbf{A}$ à une matrice de transformation géométrique définie par :

$$\mathbf{A} = \lambda \begin{bmatrix} cos\psi & -sin\psi \\ sin\psi & cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} cos\phi & -sin\phi \\ sin\phi & cos\phi \end{bmatrix} \qquad avec\ \lambda > 0, \tag{2.69}$$

il est possible de schématiser le modèle de caméra affine par la figure 2.31.

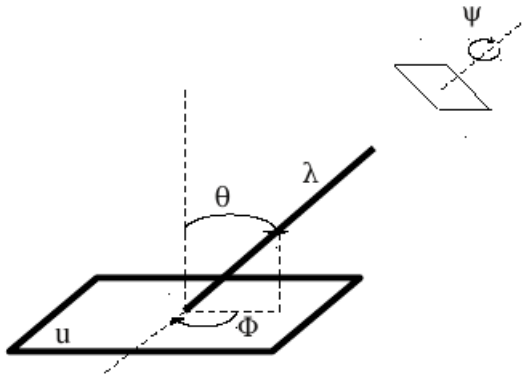


FIG. 2.31 – Modèle de caméra affine utilisé par la méthode ASIFT.
(image extraite de [86])

Les paramètres θ et ϕ déterminent l'angle de vue de la caméra par rapport au plan u. ψ représente quant à lui l'angle de rotation de la caméra. Ce modèle permet de déterminer un certain nombre d'images clés représentant l'ensemble des transformations possibles. La figure 2.32 illustre le résultat final, se résumant à la mise en correspondance de deux images A et B ainsi que de leurs transformations.

Les résultats présentés par les auteurs sont nettement supérieurs à ceux des méthodes classiques pour de fortes transformations de l'image (changement de point de vue de quatre-vingt degrés par exemple). Néanmoins le temps de calcul de cette approche augmente considérablement (de cinq à six fois plus onéreux que la méthode SIFT). Par conséquent cette méthode est utilisée dans le cas de transformations importantes de l'image n'ayant aucune contrainte temporelle.

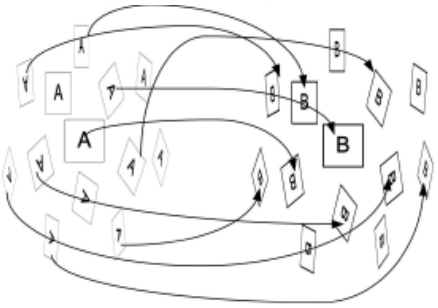


FIG. 2.32 – Mise en correspondance par la méthode ASIFT. (image extraite de [86])

B) C-HOG

a) GLOH

Détaillé en 2005 par Mikolajczyk et Schmid [83] puis testé la même année par Dalal et Triggs [32], le descripteur GLOH (Gradient location-orientation histogram) a été proposé afin d'augmenter les performances du SIFT. Il est repris en 2009 par Chandrasekhar [20] afin d'étudier toutes les possibilités de cette méthode et de l'améliorer. L'idée générale est de construire un histogramme de gradients orientés dans un plan circulaire. Pour ce faire, le GLOH est constitué de dix-sept zones d'analyse suivant trois paramètres radiaux et huit paramètres angulaires. La figure 2.33 donne un aperçu du masque d'analyse complet de ce descripteur.

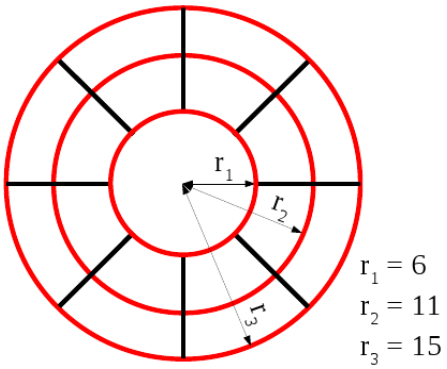


FIG. 2.33 – Masque d'analyse de la méthode GLOH, constitué de trois cercles de rayon r_1,r_2 et r_3, dont les deux plus grand sont divisés en huit zones (avec un pas de $\frac{\pi}{4}$).

En chacune des zones, un histogramme de gradients orientés est construit suivant seize classes (intervalles de $\frac{\pi}{8}$). L'histogramme final est donc constitué de 272 données, seuillées et normalisées. Les résultats obtenus prouvent que cette approche améliore dans certains cas la méthode traditionnelle du SIFT.

b) DAISY

Le descripteur DAISY proposé par Tola et al. en 2008 [107] est une nouvelle approche de description locale s'inspirant des avantages des méthodes SIFT et GLOH. Elle a pour objectif d'accélerer les temps de calculs et d'améliorer la gestion des invariances. L'idée principale est de remplacer les calculs de gradients des méthodes précédemment citées par des filtres de dérivées gaussiennes orientées. Fortement influencés par la méthode du SIFT (descripteur suivant huit classes), les auteurs proposent de créer huit orientations définies par :

$$M_\theta^\sigma = g_\sigma * (max(\frac{\partial I}{\partial \theta}, 0)), \tag{2.70}$$

où θ est l'orientation de la dérivée. Chaque M_θ ainsi créée correspond à l'ensemble des gradients d'une orientation donnée, ayant une norme positive. Pour leur descripteur Tola et al. présentent un masque d'analyse circulaire original décrit dans la figure 2.34, constitué de 25 cercles définis suivant 3 échelles.

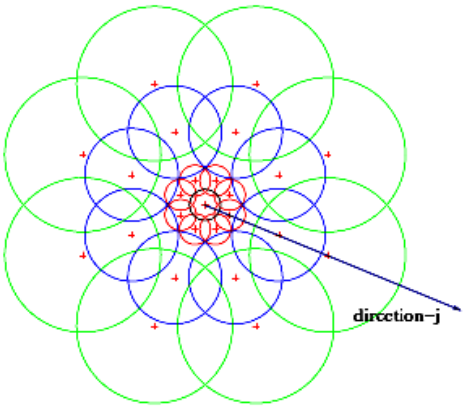


FIG. 2.34 – Masque d'analyse du descripteur Daisy, constitué de vingt-cinq cercles recalés suivant l'orientation locale du point d'intérêt. (image extraite de [107])

Dès lors, un histogramme de gradients orientés est calculé à l'intérieur de chaque cercle. En se basant sur les orientations décrites par l'équation 2.70. l'histogramme h_σ en un point $\mathbf{x}$ est défini par :

$$h_\sigma(\mathbf{x}) = [M_1^\sigma(\mathbf{x}), ..., M_8^\sigma(\mathbf{x})]^{\mathrm{T}} \tag{2.71}$$

Ce dernier est normalisé afin d'accroître l'invariance aux changements de luminosité et se note $\widetilde{h}_\sigma(\mathbf{x})$. Le descripteur final, noté $D(\mathbf{x}_0)$ se compose donc de vingt-cinq histogrammes (un par cercle) possédant chacun huit orientations et il est défini par :

$$\begin{aligned} D(\mathbf{x}_0) = [&\widetilde{h}_{\sigma_1}^{\mathrm{T}}(\mathbf{x}_0), \\ &\widetilde{h}_{\sigma_1}^{\mathrm{T}}(l_1(\mathbf{x}_0, R_1)), ..., \widetilde{h}_{\sigma_1}^{\mathrm{T}}(l_N(\mathbf{x}_0, R_1)), \\ &\widetilde{h}_{\sigma_2}^{\mathrm{T}}(l_1(\mathbf{x}_0, R_2)), ..., \widetilde{h}_{\sigma_2}^{\mathrm{T}}(l_N(\mathbf{x}_0, R_2)), \\ &\widetilde{h}_{\sigma_3}^{\mathrm{T}}(l_1(\mathbf{x}_0, R_3)), ..., \widetilde{h}_{\sigma_3}^{\mathrm{T}}(l_N(\mathbf{x}_0, R_3))]^{\mathrm{T}} \end{aligned} \tag{2.72}$$

$l_j(\mathbf{x}_0, R_i)$ représente l'indice du cercle avoisinant. Les cercles sont ordonnés suivant trois rayons : R_1, R_2 et R_3 et suivant huit orientations : $1 \rightarrow N$ avec l'indice 1 pour l'orientation locale du point d'intérêt et $N = 8$.

Au vu des résultats obtenus les auteurs conseillent les valeurs suivantes : $R_1 = 2,5$, $R_2 = 3R_1$ et $R_3 = 6R_1$ ainsi que : $\sigma_1 = 2,55$, $\sigma_2 = 3\sigma_1$ et $\sigma_3 = 5\sigma_1$. Les tests permettent également d'affirmer que la méthode Daisy offre une diminution des temps de calculs d'un facteur cinquante par rapport au SIFT.

c) Descripteur de Cheng

Proposé en 2008 par Cheng et al. [22], ce descripteur s'appuie sur une description mutli-échelles du voisinage du point d'intérêt. Après une étape d'extraction des primitives, dont le choix de la méthode de détection est libre, les auteurs créent un nombre défini de régions d'intérêt, au voisinage du point. L'image 2.35 donne un aperçu d'une telle analyse.

FIG. 2.35 – Exemple d'un descripteur de Cheng, analysant suivant s régions d'intérêt proportionnelles à l'échelle locale. (images extraites de [22])

Chaque région d'intérêt possède une taille T définie par :

$$T_s = s \times \sigma \qquad avec \;\; s = 0, ..., 2N, \tag{2.73}$$

où σ est l'échelle locale du point. La construction d'un histogramme par régions permet notamment d'accroître la robustesse aux occultations. Chaque histogramme se base sur les gradients locaux et sur leur orientation, mais à la différence des méthodes précédentes (SIFT, SURF), les auteurs préconisent de supprimer l'étape de lissage gaussien de l'image.

2.4 Les méthodes de mise en correspondance

La mise en correspondance de points d'intérêt est un processus indispensable aux applications stéréo (mettant en jeu plusieurs images). En effet il sert de passerelle entre le "haut niveau" (reconstruction, reconnaissance,...) et le "bas niveaux" (extraction

d'informations). L'objectif d'un appariement est de rechercher, dans plusieurs images, le couple de points ayant la meilleure similarité (ou ressemblance). La figure 2.36 illustre la mise en correspondance entre les ensembles de points $\{\mathbf{x}_k\}_{1\leqslant k\leqslant K}$ et $\{\mathbf{x}_l\}_{1\leqslant l\leqslant L}$, afin d'obtenir le meilleur taux d'appariement.

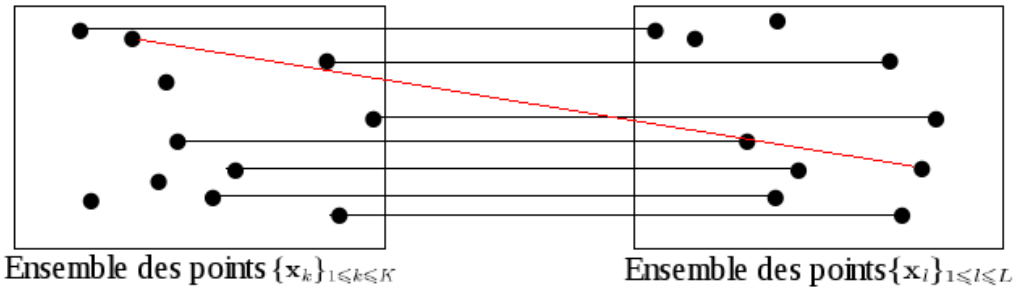


FIG. 2.36 – Représentation d'une mise en correspondance entre les points $\{\mathbf{x}_k\}_{1\leqslant k\leqslant K}$ et $\{\mathbf{x}_l\}_{1\leqslant l\leqslant L}$.

Trois cas sont alors possibles :
- les bon appariements (appelés également *inliers*) qui déterminent la qualité et la précision de mise en correspondance d'une méthode ;
- les mauvais appariements (aussi appelés *outliers*) qui détériorent les performances des applications "haut niveau", l'objectif est donc d'en diminuer le nombre ;
- les points qui ne s'apparient pas, généralement issus d'un processus cherchant à diminuer les *outliers*, ils ont l'avantage de ne pas pénaliser les applications "haut niveau".

Il existe un grand nombre de méthodes locales d'appariement de points d'intérêt et nous proposons d'en lister les différentes composantes. Ayant proposé une étude ciblée sur l'extraction et la caractérisation de points d'intérêt, notre analyse des méthodes d'appariements se limite donc à la mise en correspondance de ces derniers. Le principe est d'interpréter les informations du point d'intérêt et de son voisinage proche pour extraire le couple présentant la meilleure ressemblance. Nous proposons d'étudier les méthodes par correlation (type de mise en correspondance classique la plus utilisée), ainsi que l'appariement par relaxation et par multi-résolution (ou hiérarchie).

2.4.1 Par corrélation

Les méthodes par correlation sont principalement utilisées dans l'analyse de l'information d'intensités pour la mise en correspondance. Le principe est de déterminer, pour le voisinage d'un point $\mathbf{x}_1$ de la première image, la corrélation maximale (distance minimale) avec un voisinage issu de la seconde image. Ce calcul permet donc d'extraire le point $\mathbf{x}_2$ formant ainsi le couple ($\mathbf{x}_1$, $\mathbf{x}_2$) présentant la meilleure ressemblance au sens de la corrélation. Le schéma 2.37, résume une telle mise en correspondance.

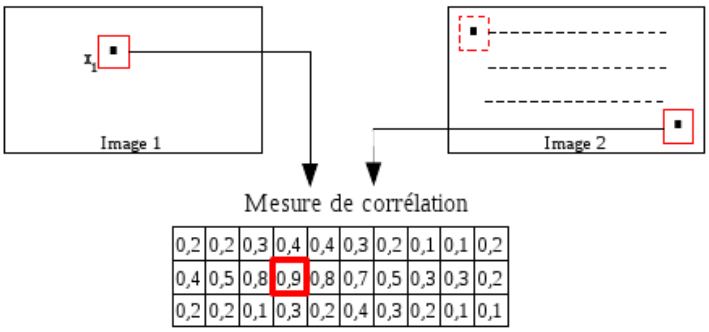


FIG. 2.37 – Principe de la corrélation : recherche du point issu de la seconde image présentant la meilleure ressemblance.

Afin d'optimiser cette méthode, une estimation de la position de $\mathbf{x}_2$ peut être introduite. Nous déterminons alors les mesures de corrélation à l'intérieur d'une zone de recherche, et non plus sur l'image entière. La figure 2.38 donne un aperçu de la zone de recherche.

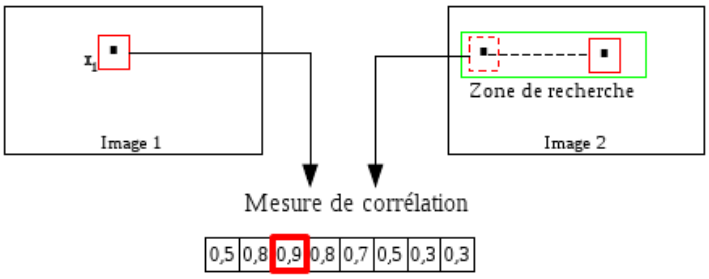


FIG. 2.38 – Exemple de mesures de corrélation entre le point $\mathbf{x}_1$ et les points $\mathbf{x}_2$ présents dans la zone de recherche.

La plus grande difficulté réside dans le choix de la mesure de corrélation. En effet il en existe un grand nombre, Aschwanden et Guggenbühl [8] proposent en 1992 de les lister et d'en donner une courte description. Les plus classiques sont la somme des distances au carré (SSD : *Sum of Squared Distances*) définie par :

$$SSD(\mathbf{x}_1;\mathbf{x}_2) = \sum_{i=-N}^{N}\sum_{j=-P}^{P}(I_1(x_1+i,y_1+j) - I_2(x_2+i,y_2+j))^2, \tag{2.74}$$

et la somme des valeurs absolues des distances (SAD : *Sum of Absolute Distances*) définie par :

$$SAD(\mathbf{x}_1;\mathbf{x}_2) = \sum_{i=-N}^{N}\sum_{j=-P}^{P}|I_1(x_1+i,y_1+j) - I_2(x_2+i,y_2+j)| \tag{2.75}$$

En ajoutant des informations locales à ces deux mesures de corrélation, il est possible d'accroître leur robustesse. D'une part, la normalisation de l'image permet de proposer la ZSSD [1] ainsi que la ZSAD [2] et d'autre part, la connaissance de l'échelle locale permet de créer la LSAD [3] et la LSSD [4]. Toujours dans une optique de robustesse, l'hypothèse d'une relation affine entre les intensités lumineuses de deux images a été proposée au travers de la ZNCC [5]. Cette dernière présente une fiabilité supérieure aux autres et est définie par :

$$ZNCC(\mathbf{x}_1;\mathbf{x}_2) = \frac{\sum_{i=-N}^{N}\sum_{j=-P}^{P}(I_1(x_1+i,y_1+j)-\overline{I_1})\cdot(I_2(x_2+i,y_2+j)-\overline{I_2})}{\sqrt{\sum_{i=-N}^{N}\sum_{j=-P}^{P}(I_1(x_1+i,y_1+j)-\overline{I_1})^2\cdot\sum_{i=-N}^{N}\sum_{j=-P}^{P}(I_2(x_2+i,y_2+j)-\overline{I_2})^2}}, \tag{2.76}$$

avec $\overline{I_k}$ représentant la moyenne de l'image I_k, déterminée par :

$$\overline{I_k} = \frac{1}{(2N+1)(2P+1)}\sum_{i=-N}^{N}\sum_{j=-P}^{P} I_k(x_k+i,y_k+j). \tag{2.77}$$

Il existe également des méthodes de corrélations utilisant un filtrage de l'image, nous pouvons citer le CC et le ZCC (respectivement *Cross-Correlation* et *Zero Cross-Correlation*). Un dernier type de mesure peut être obtenu en s'appuyant sur l'information locale de l'orientation (SES et SEK de Seitz) permettant ainsi d'accroître la robustesse aux rotations de l'image.

2.4.2 Par relaxation

Proposée par Hummel et Zucker [55] en 1983, puis améliorée par Sidibe et al. [103] en 2007, la mise en correspondance par relaxation se base sur une fonction de probabilité d'appariement. Le principe est de calculer la probabilité qu'un point $\mathbf{x}_i$ soit apparié avec un point $\mathbf{x}_j$ connaissant les appariements de ses voisins. Cette probabilité, notée $p_i(j)$ est tout d'abord initialisée, puis est mise à jour de façon itérative jusqu'à obtention d'un point stationnaire $p_i^k(j)$. La mise à jour se base sur une fonction de compatibilité q_i, définie dans le voisinage V_i du point $\mathbf{x}_i$. Il existe différents modèles d'appariement par relaxation, celui préconisé par Hummel et Zucker est défini par :

$$p_i^{k+1}(j) = \frac{p_i^k(j)q_i^k(j)}{\sum_j p_i^k(j)q_i^k(j)} \qquad avec\ q_i^k(j) = \sum_g w_{ig}\sum_h p_{ig}(j,h)p_g^k(h), \tag{2.78}$$

[1] *Zero-Mean SUM of Square Distances*
[2] *Zero-Mean SUM of Absolute Distances*
[3] *Locally scaled Sum of Absolute Distances*
[4] *Locally scaled Sum of Square Distances*
[5] *Zero-mean Normalized Cross-Correlation*

où $p_{ig}(j,h)$ est la probabilité que le point $\mathbf{x}_i$ soit apparié avec $\mathbf{x}_j$ sachant que le point $\mathbf{x}_g$ est apparié avec $\mathbf{x}_h$. Le coefficient w_{ig} permet de quantifier l'influence de $\mathbf{x}_g$ sur $\mathbf{x}_i$. Nous proposons d'illustrer ces différentes étapes, par le biais des figures 2.39,2.40 et 2.41, afin de mieux visualiser ce procédé.

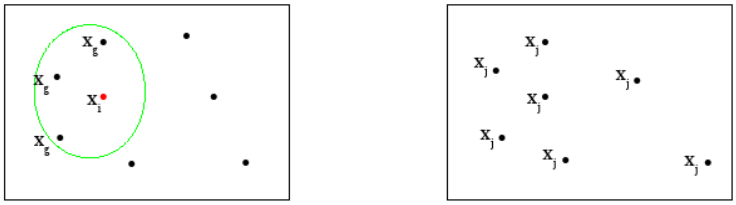


FIG. 2.39 – Etape initiale : sélection des plus proches voisins du point $\mathbf{x}_i$ et initialisation de $p_i^k(j)$

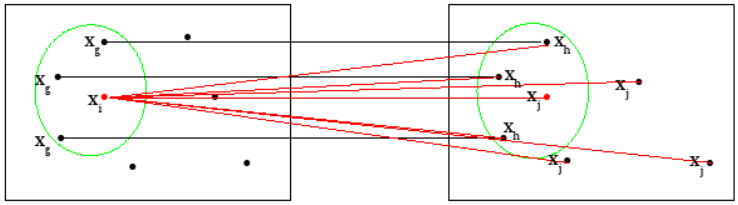


FIG. 2.40 – Etapes itératives : mise à jour de $p_i^k(j)$ en se basant sur les équations 2.78

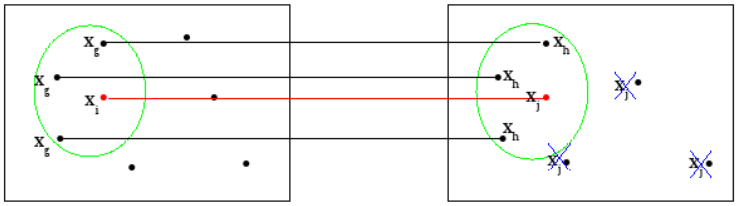


FIG. 2.41 – Etape finale : sélection de la mise en correspondance présentant la plus forte probabilité d'appariement

2.4.3 Par multi-résolution

En s'appuyant sur une mesure de corrélation de type SSD, Chen et Hung proposent en 1993 [21] une méthode multi-résolution de mise en correspondance. Cette dernière repose sur la construction d'une pyramide constituée d'images successives. L'image initiale I_l caractérise la base de la pyramide. Les étages supérieurs, $I_{l-1}...I_1$, sont calculés par lissage et échantillonnage, pour conclure par l'image I_0 répresentant le voisinage du

point d'intérêt. La valeur d'intensité du point (i, j), résultant de l'échantillonnage de I_l vers I_{l-1} est déterminée par :

$$I_{l-1}[\frac{i}{2}, \frac{j}{2}] = \frac{1}{4}(I_l[i, j] + I_l[i+1, j] + I_l[i, j+1] + I_l[i+1, j+1]) \tag{2.79}$$

Le schéma 2.42 représente la pyramide une fois construite, pour une image initiale de taille N×N.

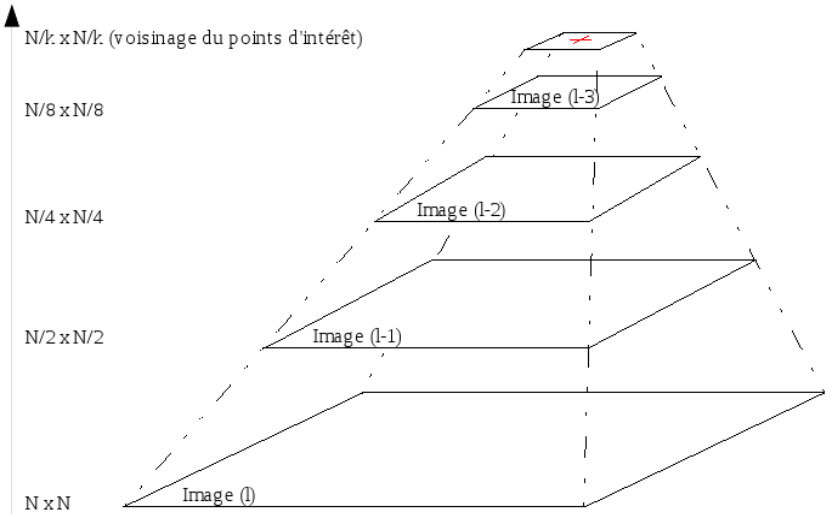


FIG. 2.42 – Exemple de pyramide constituée d'images successives, construites par échantillonnage. Leurs tailles respectives sont mentionnées sur la gauche.

Les mesures de corrélation se calculent de façon hiérarchique. Le processus débute par l'image située au sommet de la pyramide (image I_0) et se termine par l'image haute-résolution (image de départ I_l). Chaque transition $I_{l-(k+1)} \rightarrow I_{l-k}$ entraine une augmentation de la taille de la fenêtre d'analyse V du voisinage du point ainsi que celle de la zone de recherche S de la SSD.

2.5 Discussions

2.5.1 Détecteurs

L'analyse de la répétabilité de certaines méthodes classiques telles que Harris [47], DoG [31][73], hessien [81] ainsi que leur extension Harris-Laplace [80], hessien-Laplace [81] et fast-hessien [13] nous permet de mettre en avant la pertinence des points extraits et de choisir en conséquence le procédé le plus adapté à nos besoins. D'un point de vue théorique la répétabilité est définie par :

$$R = \frac{\text{Nombre de retro-projections correctes}}{\text{Nombre de points détéctés}}. \tag{2.80}$$

Afin de déterminer ce taux, une retro-projection est validée si elle répond à deux critères :

- soit $\mathbf{x}_1$ et $\mathbf{x}_2$ deux points issus de deux images distinctes, la distance les séparant après rétro-projection doit être inférieure au seuil ϵ_r (généralement égal à 1,5) :

$$\|\mathbf{x}_1 - \mathbf{H}\mathbf{x}_2\| < \epsilon_r, \tag{2.81}$$

  où $\mathbf{H}$ représente la transformée entre les deux images.

- La surface décrivant le voisinage des points doit avoir une erreur de similarité ϵ_s inférieure à un certain seuil (généralement égal à 0,4) :

$$\epsilon_s = |1 - s^2 \frac{\min(\sigma_1^2, \sigma_2^2)}{\max(\sigma_1^2, \sigma_2^2)}|, \tag{2.82}$$

  où σ_1^2 et σ_2^2 correspondent aux échelles locales des points et s caractérise la surface du voisinage.

En s'appuyant sur les détections de points d'intérêt appliquées sur diverses images nous proposons pour les méthodes précédemment citées les courbes de la figure 2.43 représentant le taux de répétabilité en fonction du type de transformation (changements de point de vue, changement d'échelle, tout deux illustrés en annexe A).

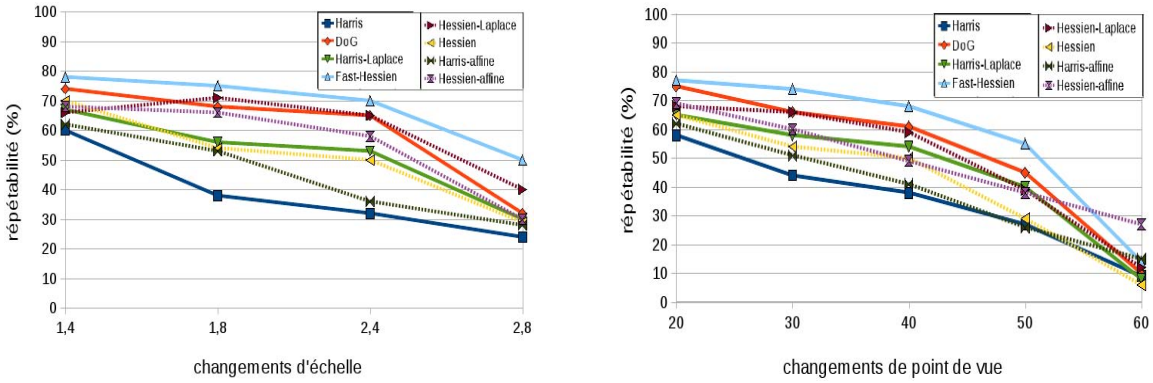


FIG. 2.43 – Courbes de répétabilité des détecteurs Harris, DoG, hessien et leurs extensions, suivant deux transformations d'images.

Pour ce type de transformations, le détecteur fast-hessien présente la meilleure répétabilité, néanmoins nous détaillerons au §3.1 certaines modifications permettant d'optimiser ses performances.

2.5.2 Descripteurs

Dans la littérature de nombreux articles proposent un référencement des méthodes existantes de description locale et les comparent les unes par rapport aux autres. Nous

pouvons citer Choksuriwong et al. [25][27], Mikolajczyk et Schmid [83][84], Bauer et al. [11] ou encore Juan et Gwun [57], regroupant ainsi une grande partie des méthodes décrites au §2.3. L'analyse des résultats présentés dans ces différents articles, concernant la précision moyenne pour des changements de point de vue (A), des changements d'échelle (B) et des changements d'illumination (C), permet d'établir le tableau suivant :

	HOG			Histogramme de couleur	Moment Zernike	Transformée de Fourier
	SURF	SIFT	DAISY			
A	83,88%	79,98%	**85,34%**	59,44%	66,31%	67,50%
B	**94,78%**	93,36%	94,01%	75,65%	78,08%	81,76%
C	96,46%	**96,59%**	94,78%	82,13%	93,14%	79,44%

Il apparait que l'utilisation d'histogramme de gradients orientés (HOG) est l'approche présentant les meilleurs résultats. Cette dernière reste également, à l'heure actuelle, la plus utilisée pour la construction de descripteurs locaux. Nous avons également pu différencier deux types de HOG utilisant soit un masque à forme carrée (R-HOG) soit un masque circulaire (C-HOG). Le premier type d'histogramme est notamment utilisé par les méthodes SIFT et SURF, le second par les méthodes Daisy et GLOH. En s'appuyant sur les différentes caractéristiques des HOG nous proposons d'utiliser un masque elliptique créant ainsi des E-HOG (HOG à base elliptique). Les différents paramètres nécessaires à la construction de ce masque sont détaillés au §3.2.

2.5.3 Mises en correspondance

Nous avons présenté au §2.4 diverses méthodes de mise en correspondance. Il en résulte que l'approche par corrélation est la plus populaire et présente une mise en oeuvre ne nécessitant aucune connaissance préalable (concernant l'image, les formes ou autres). Elle peut être utilisée pour des appariements basés sur des coordonnées, des valeurs d'intensités, des régions d'intérêt ou, dans notre cas, des descripteurs locaux. En s'appuyant sur les données issues des histogrammes de gradients orientés, il est donc possible de définir la distance euclidienne inter-descripteur d_e par :

$$d_e(\mathbf{f}_{I_1}(\mathbf{x}_0), \mathbf{f}_{I_2}(\mathbf{x}_l)) = \sqrt{(\mathbf{f}_{I_1}(\mathbf{x}_0) - \mathbf{f}_{I_2}(\mathbf{x}_l))^{\mathrm{T}} \cdot (\mathbf{f}_{I_1}(\mathbf{x}_0) - \mathbf{f}_{I_2}(\mathbf{x}_l))}, \tag{2.83}$$

où $\mathbf{f}_{I_n}(\mathbf{x}_k)$ correspond au descripteur du point $\mathbf{x}_k$ issu de l'image I_n et p représente sa taille. Pour apparier le point $\mathbf{x}_0$ avec un point de la liste $\{\mathbf{x}_l\}_{0 \leqslant l \leqslant L-1}$, la distance d_e est minimisée :

$$\tilde{l} = \underset{l \in [\![0;L-1]\!]}{\operatorname{argmin}} (d_e(\mathbf{f}_{I_1}(\mathbf{x}_0), \mathbf{f}_{I_2}(\mathbf{x}_l))). \tag{2.84}$$

Cette minimisation permet ainsi d'obtenir le couple de points $\{\mathbf{x}_0; \mathbf{x}_{\tilde{l}}\}$ présentant un maximum de corrélation (voir figure 2.37). Nous proposons au §3.3 l'utilisation d'un arbre de décision permettant d'extraire rapidement cette mesure de corrélation. Nous ajoutons également un seuil de validation et une méthode de suppression des doublons.

2.5.4 Conclusion

En s'appuyant sur ces différents constats, nous allons détailler au chapitre suivant la construction de notre méthode. Elle s'appuie sur :

- une version optimisée du détecteur fast-hessien ;
- l'utilisation d'un nouveau masque d'analyse adaptatif permettant la description locale du voisinage ;
- une mise en correspondance restrictive.

L'objectif est de fournir des couples de points présentant la meilleure précision possible, permettant ainsi d'améliorer les performances des différentes applications visées.

3 Méthode REFA : Analyse locale de points d'intérêt pour des appariements robustes

Un problème récurrent dans le domaine du traitement d'image concerne l'invariance aux éventuelles transformations que cette dernière peut subir (rotation, changement d'échelle et changement de point de vue). Afin d'être le plus robuste possible, nous proposons dans ce chapitre une approche, nommée REFA (*Robust E-hog for Features Analysis*), se basant à la fois sur des outils existant ainsi que sur des optimisations et sur un masque original. Une étude comparative des approches (détecteur, descripteur et mise en correspondance) détaillées au chapitre précédent nous permet de sélectionner ceux répondant le mieux à notre problématique. Dans un premier temps nous justifierons notre choix concernant la méthode de détection fast-hessien et nous détaillerons les modifications apportées. Par la suite nous dévelloperons notre méthode de description locale s'appuyant sur un masque elliptique permettant de calculer des histogrammes de gradients orientés (E-HOG). Nous détaillerons également la méthode de mise en correspondance choisie, basée sur un calcul de corrélation et sur des outils permettant de l'optimiser. Afin de valider notre approche, nous proposerons un ensemble de tests et résultats s'appuyant sur des images synthétiques et des images réelles. Nous proposerons également quelques résultats concernant l'estimation de la matrice d'homographie et l'utilisation de notre méthode sur des séquences d'images.

3.1 Choix et optimisation de la méthode de détection

La première étape consiste à extraire des primitives présentant des données importantes de l'image (les coins, les contours par exemple). Nous avons proposé au §2.5 une comparaison de différentes méthodes existantes afin de déterminer celles présentant le meilleur taux de répétabilité. En effet, cette caractéristique met en avant le nombre de points d'intérêt extraits se répétant d'une image sur l'autre, influant directement sur les performances du descripteur. Il en résulte que le détecteur fast-hessien, notre choix, présente le meilleur taux de répétabilité pour les transformations étudiées. Il permet donc d'obtenir un taux de mises en correspondance conséquent.

Le détecteur fast-hessien détaillé au §2.2.2.3 se base donc sur l'interprétation de la matrice hessienne 2.45, dont le déterminant se calcule de la façon suivante :

$$\det(\mathbf{H}(\mathbf{x};\sigma)) = \sigma^2(L_{xx}(\mathbf{x};\sigma)L_{yy}(\mathbf{x};\sigma) - L_{xy}^2(\mathbf{x};\sigma)). \tag{3.1}$$

La recherche de maxima locaux de ce déterminant nous permet d'établir une liste de K points associés à une échelle caractéristique et définie par :

$$(\mathbf{x}_k;\sigma_k) = \underset{\{\mathbf{x};\sigma\}}{\text{argmax}}(\det(\mathbf{H}(\mathbf{x};\sigma))) \qquad avec\ k \in [\![0;K-1]\!]. \tag{3.2}$$

Afin d'optimiser cette détection, nous apportons certaines modifications à cet outil. Ces améliorations concernent d'une part l'espace d'échelle de recherche et d'autre part un seuillage permettant de sélectionner uniquement les points présentant le meilleur score de détection.

Initialement le détecteur s'appuie sur une recherche multi-échelle répartie en quatre octaves. Chacun d'entre eux est constitué de quatre filtres d'approximation présentés en figure 2.18. La taille de ces derniers est doublée pour chaque nouvel octave. Cet aspect peut être représenté de la façon suivante :

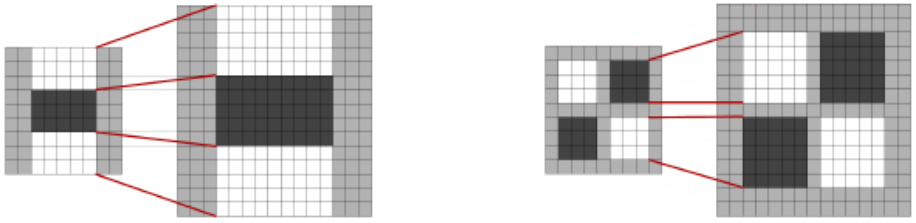


FIG. 3.1 – Représentation d'un changement d'octave (images extraites de [13]).

Par conséquence, nous proposons d'analyser l'influence du nombre d'octaves d'exploration sur les performances de notre méthode pour une transformation de type changement de point de vue (figure 3.2).

FIG. 3.2 – Exemple de transformations de type changement de point de vue (Graffiti). De gauche à droite, valeur de l'angle d'orientation de la caméra suivant un plan normal à l'image : 0°, 20°, 30°, 40° et 50°.

La figure 3.3 résume les résultats ainsi obtenus, en terme de taux de répétabilité (pour les 500 meilleurs points de chaque détecteurs) et de nombre de points détectés.

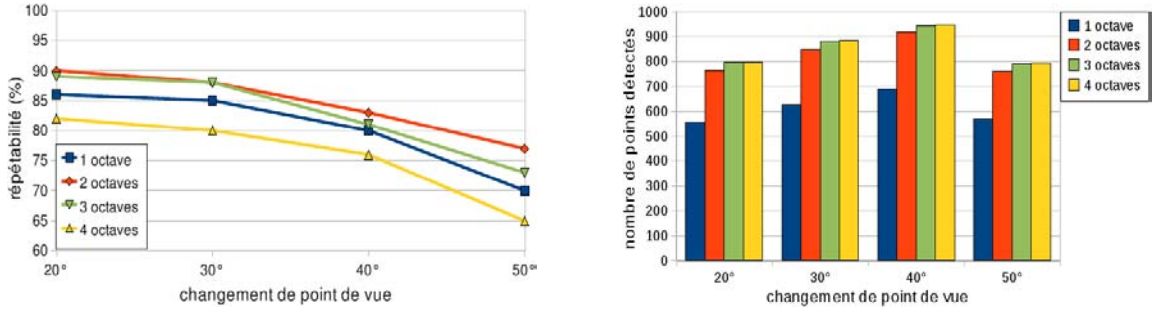


FIG. 3.3 – Influence du nombre d'octaves sur le taux de répétabilité (seuillage des 500 meilleurs points) et sur le nombre de points détéctés (sans seuillage), pour une transformation de type changement de point de vue (Graffiti).

Au vu des résultats présentés, nous pouvons affirmer qu'une diminution du nombre d'octaves entraîne un accroissement de la répétabilité du détecteur au dépend du nombre de points détectés. Par conséquent, nous avons opté pour une analyse multi-échelle suivant deux octaves (figure 3.3). En effet, en acceptant une perte moyenne de $2,81\%$ des points détéctés (soit 96 points supprimés sur les 3415 initialement détectés), notre approche obtient une augmentation de ces performances pouvant s'élever à $12,85\%$ (changement de point de vue de 50°) en terme de répétabilité.

Nous avons également étudier la "qualité" des points extraits, se basant sur le score de détection retourné par le détecteur. En seuillant ce dernier, nous avons pu observer son influence sur les performances de notre méthode. La figure 3.4 illustre cette optimisation pour une transformation de type changement de point de vue.

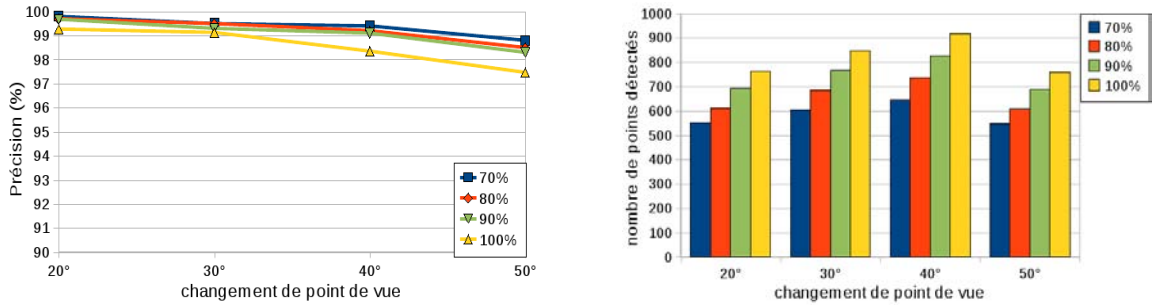


FIG. 3.4 – Influence du score de détection sur notre méthode pour une transformation de type changement de point de vue (Graffiti).

Les graphiques obtenus montrent qu'une suppression des points, ceux présentant les

scores de détection les plus faibles, permet d'accroître nos résultats. Afin de conserver le plus grand nombre de points possible, notre choix de seuil à 90% de sélection reste le meilleur compromis entre performance et nombre de points appariés.

En définitive notre méthode de détection se base sur le détecteur fast-hessien car il présente le meilleur taux de répétabilité. Nous proposons également deux optimisations, la première limitant l'espace d'échelle de recherche à deux octaves et la seconde permettant de supprimer 10% des points ayant les scores de détection les plus faibles.

3.2 Descripteur REFA

En s'appuyant sur les points forts des HOG et en gardant pour objectif d'obtenir la plus grande invariance possible, nous proposons d'utiliser un masque elliptique afin de rester le plus fidèle possible à la distribution locale de l'information.

3.2.1 Masque d'analyse

En s'appuyant sur les différents articles de la littérature précédemment cités, il convient qu'un masque à base circulaire permet de mieux définir le voisinage d'un point d'intérêt. En effet, Tola et al. [107] et Mickolajczyk et Schmid [83] détaillent les différents avantages à utiliser un tel masque. D'une part la description circulaire du voisinage d'un point est plus pertinente car elle permet de récupérer la même quantité d'informations quelque soit l'orientation et de la pondérer de façon uniforme. D'autre part l'utilisation de C-HOG permet, comme dans le cas du descripteur Daisy [107], de pouvoir combiner de façon identique les masques circulaires. La figure 3.5 illustre la différence entre deux masques d'analyses : l'un composé de neuf cercles et l'autre de neuf carrés.

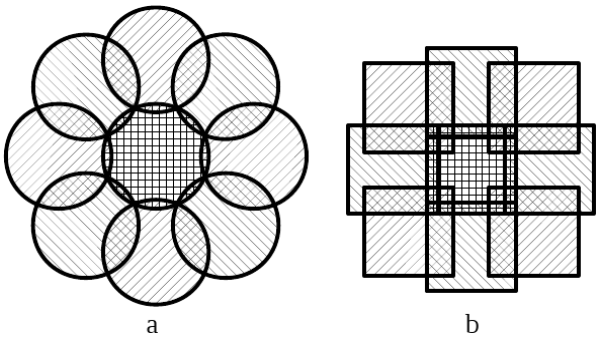


FIG. 3.5 – Exemple de masques d'analyse : à gauche un masque constitué de neuf cercles et à droite un masque constitué de neuf carrés.

Le premier constat est que chaque cercle constituant le premier masque (figure 3.5.a) décrit une région unique et de taille identique (hachure simple). Un second avantage réside dans la minimisation de la superposition des données notamment entre le

cercle central et les cercles periphériques est également minimisée. Le dernier avantage concerne l'angle de recalage détaillé au §3.2.2. Ce dernier a une plus forte influence sur le second masque du fait qu'une erreur de quelques degrés modifie la description locale. Cette perturbation est moins importante dans le cas d'une description circulaire.

En se basant sur ces différentes observations, nous avons proposé dans un premier temps un masque d'analyse constitué de dix-sept cercles. La figure 3.6 représente notre masque de description, centré sur le point d'intérêt.

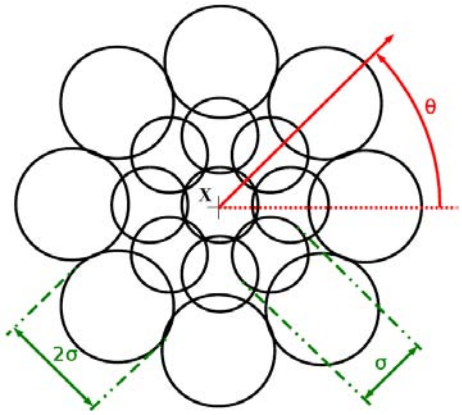


FIG. 3.6 – Représentation de notre masque d'analyse initial.

Le paramètre θ sera défini aux §3.2.2 et le paramètre σ est extrait par le biais du détecteur. D'un point de vu local, une transformation affine peut être approximée par un changement d'échelle anisotrope. Par conséquent, il devient judicieux de modifier notre masque afin d'obtenir une meilleure description du voisinage. Nous proposons d'appliquer une déformation anisotropique au masque initial créant ainsi des ellipses, dont la description locale s'appuie sur des E-HOG (généralisation anisotropique des C-HOG). Nous obtenons le masque suivant :

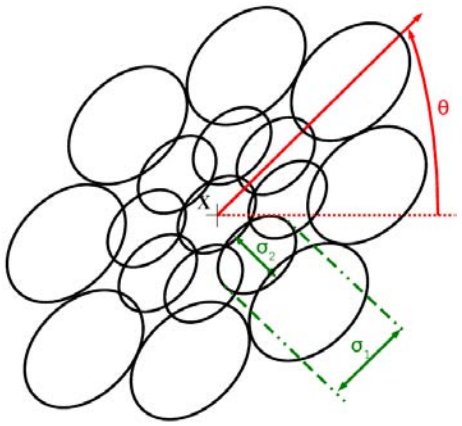


FIG. 3.7 – Représentation de notre masque d'analyse final.

Les paramètres σ_1 et σ_2 représentant respectivement le grand axe et le petit axe des ellipses sont extraits de la matrice de Harris. Nous en détaillerons le calcul au §3.2.3.

3.2.2 Ajustement de l'analyse locale : recalage du masque

Afin d'être le plus robuste possible aux rotations, l'estimation de l'orientation locale du gradient du point d'intérêt est nécessaire. Ce paramètre permet de recaler les HOG, ce qui correspond à les orienter de façon identique pour deux points correspondants. Pour cela, nous allons utiliser la matrice de Harris $\mathbf{M}$ (équation 2.17) calculée dans un voisinage circulaire de rayon σ (σ étant l'échelle locale retournée par le détecteur), centré en chaque point $\mathbf{x}$. Les dérivées premières de l'image sont déterminées à l'aide de l'opérateur de Canny-Deriche. Les propriétés de cette matrice permettent notamment d'étudier la dispersion de l'information. L'analyse locale du vecteur propre $\overrightarrow{\mathbf{v}_1}$ associé à la plus forte valeur propre de la matrice $\mathbf{M}$ permet d'extraire une estimation de l'orientation :

$$\hat{\theta}_k = \arctan(\overrightarrow{\mathbf{v}_1}). \tag{3.3}$$

3.2.3 Détermination des échelles σ_1 et $\sigma 2$

Cette étape consiste à s'appuyer sur les valeurs propres (λ_{max} et λ_{min}) issues de la matrice de Harris $\mathbf{M}$ (équation 2.17) afin de déterminer la taille de chaque ellipse. Comme illustrer en figure 3.7, deux paramètres sont nécessaire : σ_1 et σ_2. Dans la littérature, la demi-longueur du grand axe ($\sigma_1/2$) est définie comme étant l'inverse de la racine carrée de la plus petite valeur propre. Afin de conserver une cohérence dans l'analyse de l'information, le rapport r_0 entre σ_1 et σ_2 doit être compris entre $0,5$ et 1. En effet un rapport inférieur à $0,5$ "écraserait" l'éllipse, ramenant l'analyse à un simple segment, ce qui annulerait l'intérêt d'un masque adaptatif. Nous proposons donc de définir r_0 de la façon suivante :

$$r_0 = \begin{cases} 0,5 \text{ si } r < 0,5 & \text{avec } r = \frac{\sqrt{\lambda_{min}(\mathbf{M}_H)}}{\sqrt{\lambda_{max}(\mathbf{M}_H)}}. \\ r \text{ sinon} \end{cases} \tag{3.4}$$

En définitive, $\sigma_1 = \frac{2}{r_0\sqrt{\lambda_{max}(\mathbf{M}_H)}}$ et $\sigma_2 = \frac{2}{\sqrt{\lambda_{max}(\mathbf{M}_H)}}$. Le seuillage par l'intervalle $[\![0,5;1]\!]$ permet donc de conserver un rapport cohérent entre ces deux échelles.

3.2.4 Construction du descripteur

La construction de notre descripteur s'appuie sur la norme du gradient et sur son orientation pour chaque pixel inclu dans l'une des dix-sept ellipses. L'orientation du gradient nous permet de déterminer la classe à laquelle il appartient. En nous basant sur les courbes et le diagramme de la figure 3.8, présentant la précision et le nombre de points appariés pour les images Grafiti, Leuven et Boat (figure 3.9), nous optons pour l'utilisation d'histogrammes composés de huit classes et créons ainsi un descripteur appartenant à $\mathbb{R}^{136}$ (17 histogrammes de 8 classes chacuns).

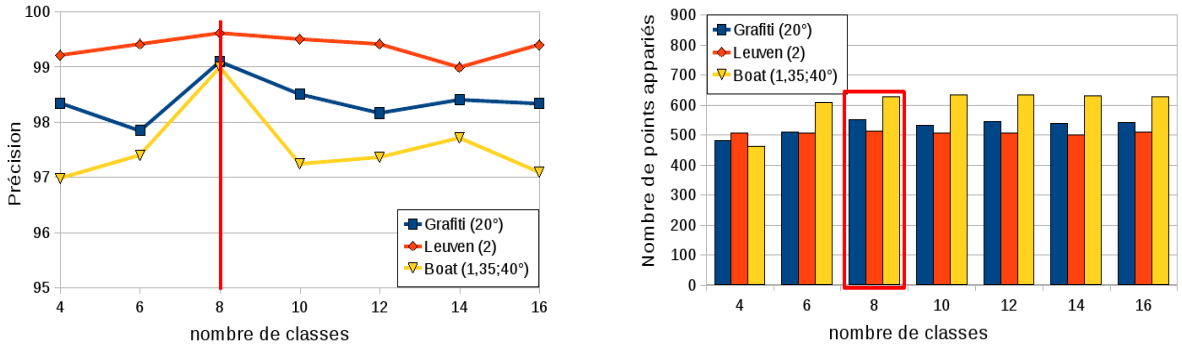


FIG. 3.8 – Précision et nombre d'appariement obtenus pour les images Graffiti, Leuven et Boat, utilisés pour déterminer le nombre de classes des histogrammes.

a b c

FIG. 3.9 – Images (a) Graffiti, (b) Leuven et (c) Boat présentant respectivement des transformations de type changement de point de vue, changement de luminosité et couplage changement d'échelle/rotation.

La dernière étape de notre méthode de description consiste d'une part, à atténuer l'influence des forts gradients et d'autre part, à normaliser les histogrammes ainsi obtenus. Dans les cas de forts gradients et conserve, leurs saturations permettent de conserver le reste de l'information contenue dans le voisinage. Ce critère est déterminé avec des tests pour différents types de transformations (voir figure 3.10). Il apparaît que le choix d'un seuil de saturation égal à $0,5$ présente les meilleurs performances en terme de précision. La normalisation des histogrammes par rapport à la norme L_∞ nous permet quant à elle d'accroître la robustesse aux éventuels changements de luminosité.

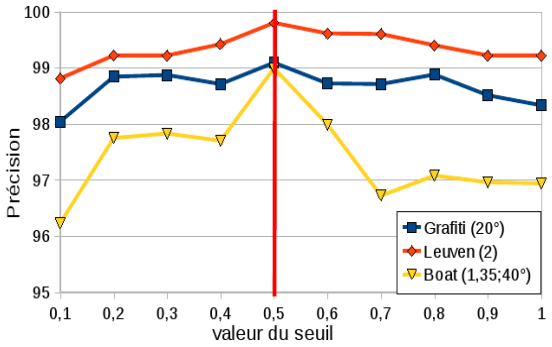


FIG. 3.10 – Détermination du seuil de saturation des histogrammes de gradients.

Notre analyse du voisinage des points d'intérêt se résume par la construction de dix-sept HOG suivant dix-sept ellipses recalées. Ces dernières permettent de récupérer l'information locale de façon pertinente. Les histogrammes de gradients orientés ainsi obtenus sont saturés puis normalisés. Une telle caractérisation est généralement suivie d'une mise en correspondance afin d'extraire des couples de points d'intérêt pour des applications de "haut niveau".

3.3 Mise en correspondance

L'étape de mise en correspondance s'appuie sur la minimisation des distances inter-descripteurs. Nous ne souhaitons introduire aucune donnée supplémentaire à notre calcul de corrélation (estimation de la position, informations sur les descripteurs avoisinants par exemple). De ce fait, pour chaque descripteur l'ensemble des candidats est testé, afin d'en extraire celui présentant la plus forte ressemblance (méthodes des plus proches voisins). Pour se faire nous proposons d'utiliser, d'une part, une méthode d'approximation de l'extraction des plus proches voisins (type ann-k) afin notamment de diminuer les temps de calculs, et d'autre part, une optimisation en terme de validation et d'unicité.

3.3.1 Construction de l'arbre de décision

La méthode d'extraction des k plus proches voisins, nommée ann-k (*Approximate Nearest Neighbor search method*) et développée de 1993 à 1994 par Arya et al. [5][6], repose sur la construction d'un arbre de décision. Les auteurs s'appuient notamment sur les travaux de Friedman et al. [43] qui proposent en 1977 une méthode de construction optimisée. Le principe est de subdiviser un modèle plan de façon itérative par des hyperplan afin de créer un référentiel. De nombreuses optimisations ont été proposées : Clarkson en 1983 [29] et 1994 [30], Bern at al. en 1993 [17] ainsi que Arya et al. [5][6]. La construction de l'arbre de décision et l'extraction des k plus proches voisins sont schématisées par la figure 3.11.

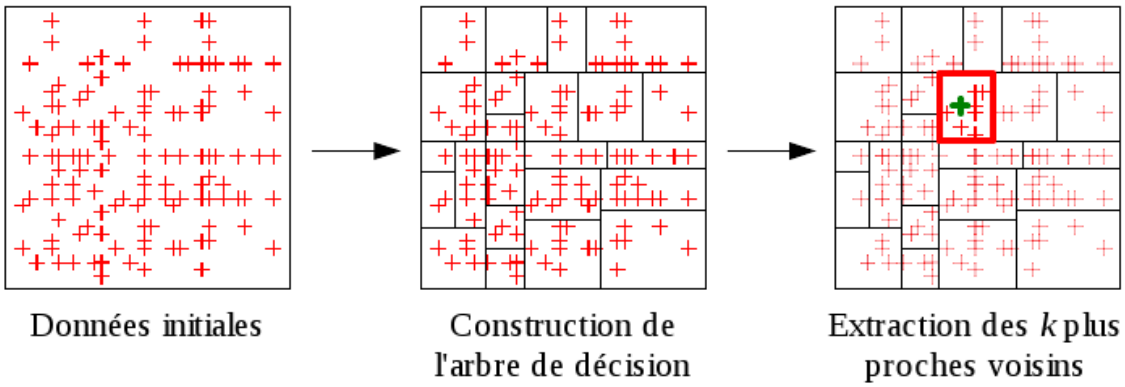


FIG. 3.11 – Arbre de décision : construction et extraction des k plus proches voisins (par rapport au point vert) parmi l'ensemble des candidats (points rouges) formant le référentiel.

Les données initiales, dans notre cas les descripteurs, permettent de former le référentiel afin de déterminer parmi tous les candidats, ceux présentant la meilleure ressemblance avec le point à apparier. La dimension de cet arbre dépend de la taille des données que nous lui fournissons. L'utilisant pour extraire les meilleurs candidats pour notre mise en correspondance l'arbre de décision est donc de dimension 136, correspondant à la taille de chacun de nos descripteurs.

3.3.2 Optimisation des appariements

Afin de rendre plus robuste la mise en correspondance, il est possible d'utiliser un seuil de sélection. Ce dernier permet d'accentuer la pertinence des appariements en n'acceptant que les couples de points présentant un score de ressemblance élevé. Le principe est d'extraire tout d'abord les k plus proches voisins (dans notre cas les deux plus proches voisins) puis de comparer les scores d'appariement obtenus pour chacun d'entre eux. Nous proposons pour notre approche, en s'appuyant sur l'équation 2.84, de valider une mise en correspondance si et seulement si :

$$d_e(\mathbf{f}_{I_1}(x_0, y_0), \mathbf{f}_{I_2}(x_{\tilde{l}}, y_{\tilde{l}})) \leq \alpha \times min(d_e(\mathbf{f}_{I_1}(x_0, y_0), \mathbf{f}_{I_2}(x_l, y_l))), \tag{3.5}$$

où $l \in [\![0; N-1]\!] \backslash \tilde{l}$ et α correspond au seuil de sélection. Afin de déterminer expérimentalement le coefficient α, une étude de son influence sur les performances de notre méthode est présentés sur la figure 3.12.

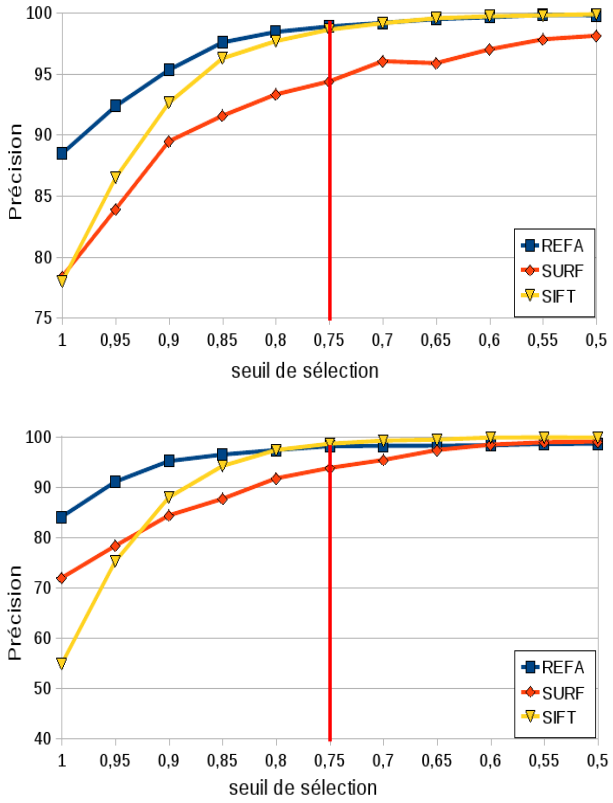


FIG. 3.12 – Détermination du seuil de sélection pour un changement de point de vue de 20° (haut) et un couplage rotation/changement d'échelle de $40°/1,35$ (bas).

Au vu des résultats, il apparaît que la précision de notre méthode devient constante pour une valeur de seuil égale à 0,75. De plus ce choix correspond à un bon compromis entre la précision et le nombre de points mis en correspondance. En effet, la diminution de cette valeur de seuil entraîne une perte de précision. Son augmentation implique quant à elle une plus grande sélectivité et donc un nombre plus faible d'appariements. Pour l'ensemble des résultats présentés au §3.5, nous optons donc pour un coefficient de sélection α égal à 0,75.

Nous proposons une dernière étape d'optimisation de la mise en correspondance reposant sur une suppression des doublons. Le principe réside dans le fait que nous souhaitons interdire à un point d'intérêt de s'apparier à plusieurs candidats. Pour ce faire les appariements sont classés et les points présentant plus d'une mise en correspondance sont supprimés. Le seul inconvénient de ce procédé est la suppression à la fois de bons et de mauvais appariements. Néanmoins nous préférons éliminer quelques appariements corrects, permettant également de supprimer des incorrects et par conséquent d'accroître la précision.

3.4 Résumé de la méthode REFA

Afin d'avoir une vue d'ensemble des différentes étapes nécessaires à la construction de notre approche, nous proposons de les schématiser (figure 3.13) et d'en donner une courte description. Dans cette illustration, nous mettons en correspondance deux images dont l'une a subit une rotation de 45°. Pour assurer une meilleure visibilité, seule une partie des points à analyser est représentée.

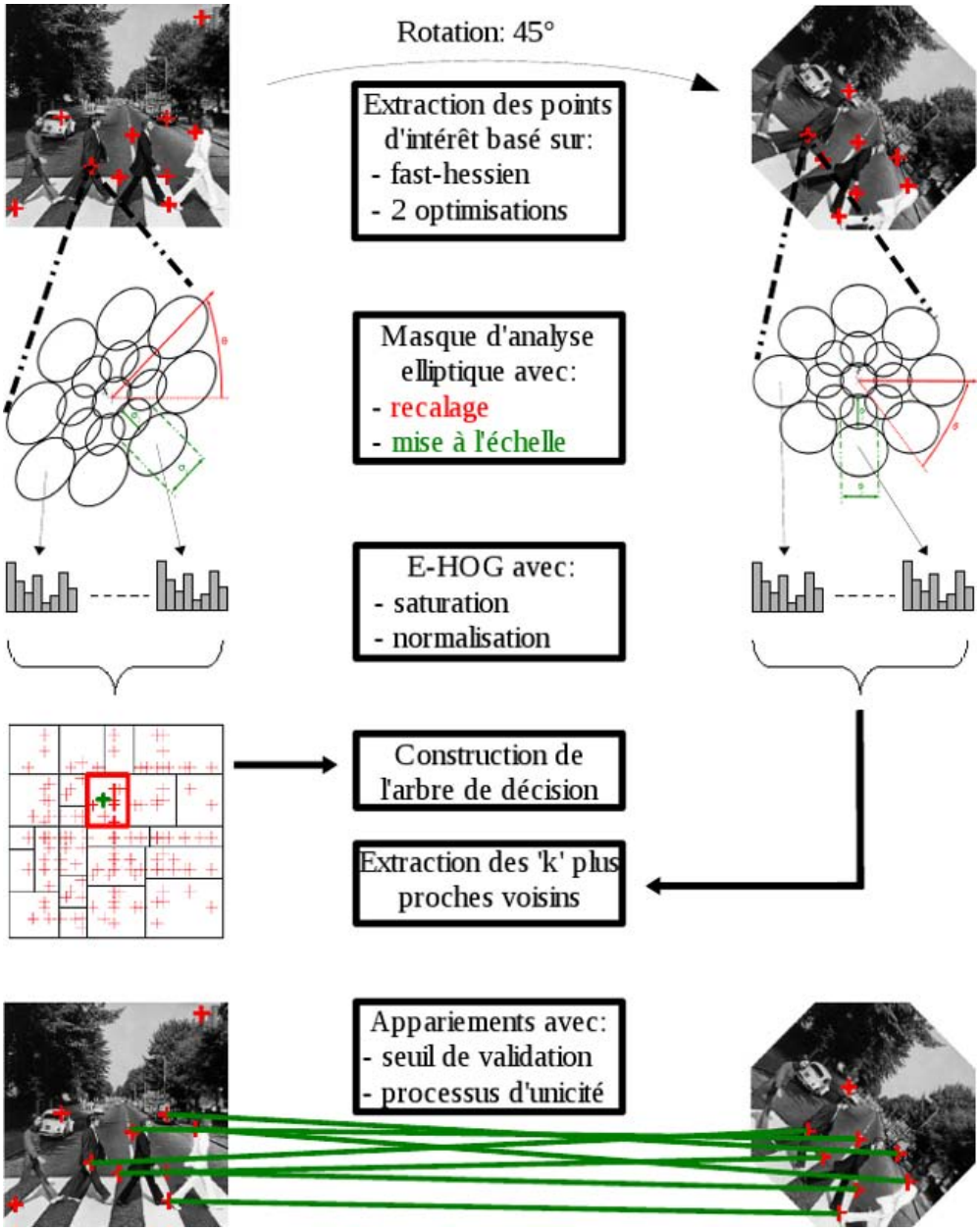


FIG. 3.13 – Schéma récapitulatif décrivant les différentes étapes de construction de la méthode REFA.

3.5 Validation de notre méthode

Les différents tests de validation permettent de mettre en avant les avantages et les inconvénients de la méthode REFA. Pour ce faire nous utilisons un nombre conséquent d'images, présentées en Annexe A, regroupant ainsi les différents types de transformations possibles. Nous proposons d'étudier quatre critères d'observation définis comme suit :

- la **précision**, définie par :

$$\text{Précision} = \frac{\text{Nombre d'appariement correct}(I_1, I_2)}{\text{Nombre de correspondance trouvée}(I_1, I_2)} \tag{3.6}$$

  Ce taux permet d'évaluer la qualité de la mise en correspondance, et à fortiori la pertinence de notre description locale.

- le **nombre de points appariés** et le **taux d'appariement**. Le premier caractérise l'aspect quantitatif de l'appariement et le second est défini par :

$$\text{Taux d'appariement} = \frac{\text{Nombre de points appariés}(I_1, I_2)}{\text{Nombre de correspondance possible}(I_1, I_2)} \tag{3.7}$$

  Ce taux permet de valoriser les performances du descripteur en fonction du nombre de points extraits par le détecteur.

- un critère reposant sur une analyse couramment utilisée : ***recall* en fonction de 1-Précision**, dont ses composantes sont définies par :

$$\textit{recall} = \frac{\text{Nombre d'appariement correct}(I_1, I_2)}{\text{Nombre de correspondance correct possible}(I_1, I_2)} \tag{3.8}$$

  et

$$1 - \text{Précision} = \frac{\text{Nombre d'appariement incorrect}(I_1, I_2)}{\text{Nombre de correspondance trouvée}(I_1, I_2)} \tag{3.9}$$

  Cette analyse permet d'étudier l'influence de la détérioration des données sur la précision et la qualité de la mise en correspondance.

L'ensemble de ces critères permet de vérifier que les objectifs annoncés sont validés : la qualité et la pertinence de notre description locale, la mise en correspondance d'un nombre correct de points d'intérêt et enfin la robustesse de notre approche.

3.5.1 Comparaison avec les méthodes SIFT et SURF

Au cours du chapitre précédent, nous avons détaillé un certain nombre de méthodes de détection et de description locales. Souhaitant valider notre approche, nous proposons de la comparer à deux des méthodes les plus utilisées et les plus performantes. En effet

la littérature [11][25][57][83] désigne les méthodes SIFT et SURF comme celles ayant les meilleurs résultats en termes de taux d'appariement et de robustesse aux différentes transformations de l'image. Nous proposons d'observer les différents critères énoncés précédemment pour des transformations synthétiques puis réelles. Nous analyserons également les résultats obtenus concernant l'estimation de la matrice d'homographie.

3.5.2 Transformations synthétiques : comparaison à la vérité terrain

Nous proposons d'analyser dans un premier temps des transformations uniques telles que le changement d'échelle, l'étirement et la rotation. Pour l'ensemble des résultats, nous présentons les courbes de précision et les diagrammes décrivant le nombre d'appariement effectué. Pour une meilleure lecture nous illustrons la transformation étudiée par un échantillon des images utilisées.

– Concernant les transformations de type **changements d'échelles isotropes**, les figures F.1,F.2 et F.3 présentent les résultats obtenus pour les images Beatles, Lena et Leuven (voir Annexe A).

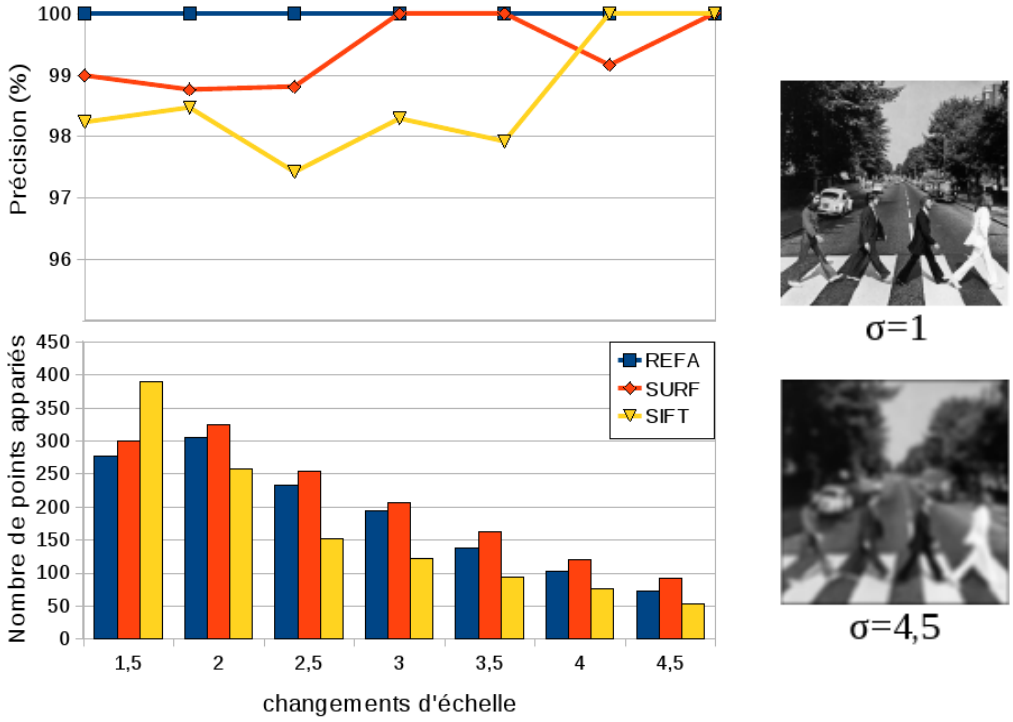


FIG. 3.14 – Précision et nombre de points appariés pour des changements d'échelle sur les images Beatles.

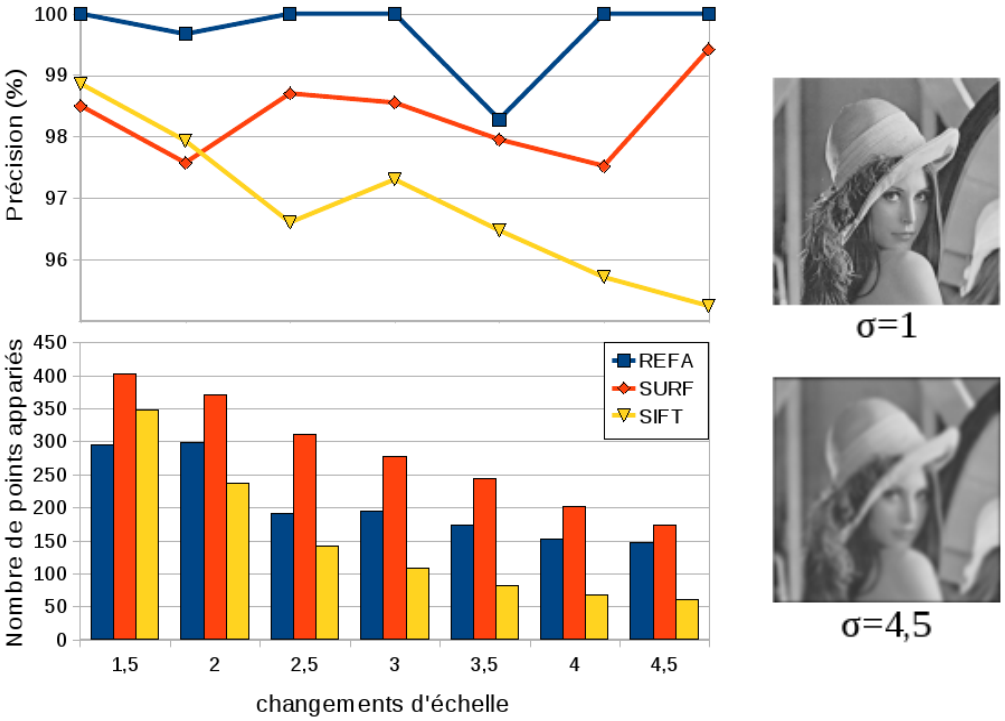


FIG. 3.15 – Précision et nombre de points appariés pour des changements d'échelle sur les images Lena.

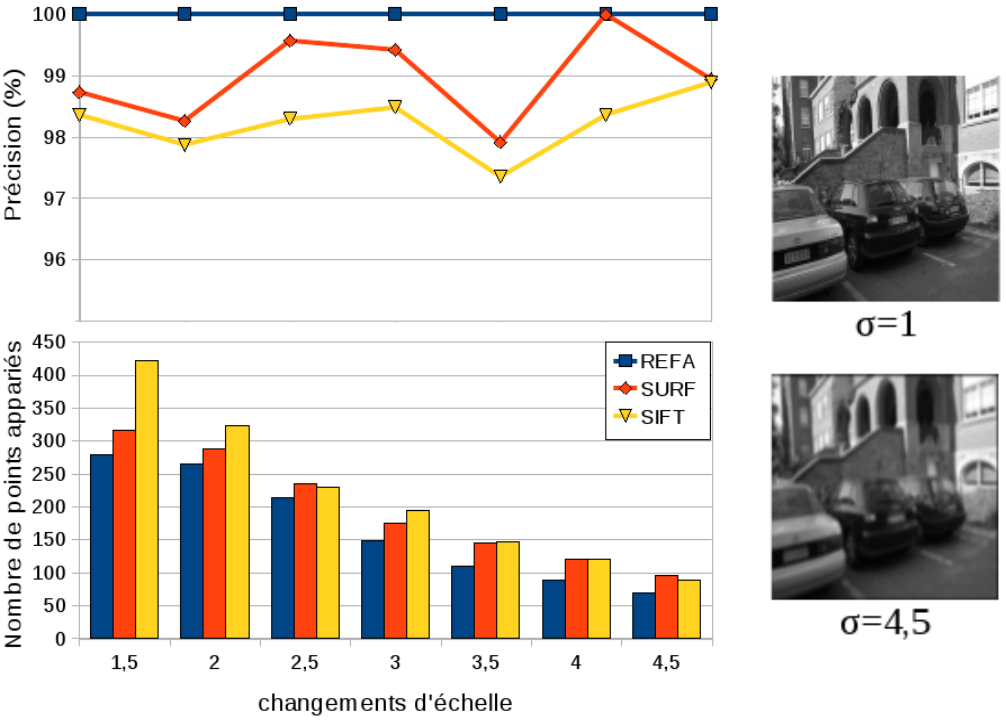


FIG. 3.16 – Précision et nombre de points appariés pour des changements d'échelle sur les images Leuven.

Pour ce type de transformation, notre approche présente une meilleure précision que le SIFT et le SURF au dépend d'un nombre d'appariements plus faible, notamment pour les premières valeurs d'échelles. Nous verrons tout au long des tests effectués que l'utilisation d'ellipses comme région d'analyse permet d'accroître les performances de notre approche. Dans le cas de transformations de type changements d'échelles, les paramètres σ_1 et σ_2 (définis sur l'image 3.7) jouent un rôle prépondérant. En effet la mise à l'échelle de notre masque d'analyse permet d'extraire l'information de façon identique d'une image sur l'autre. Cette adaptation explique donc les résultats pouvant atteindre 100% en terme de précision pour des données synthétiques.

– Nous proposons d'analyser des transformations de type **étirements unidirectionnels**. Les figures F.4,F.5 et F.6 regroupent les résultats obtenus pour les images Lena, Pig et Graffiti.

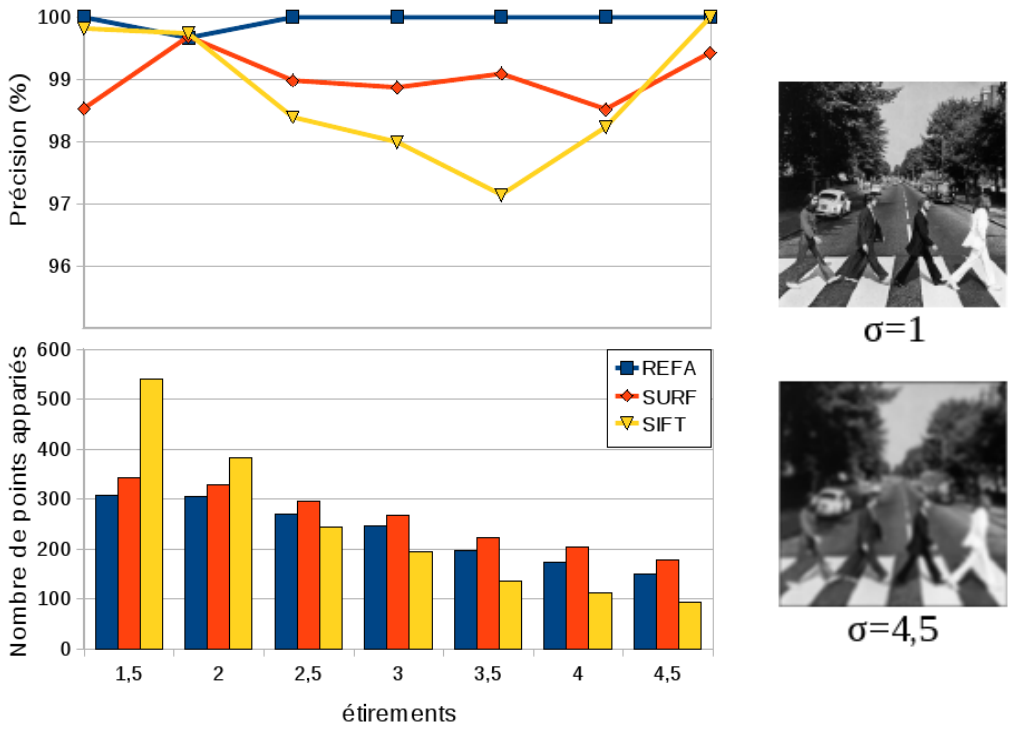


FIG. 3.17 – Précision et nombre de points appariés pour des étirements unidirectionnels sur les images Beatles.

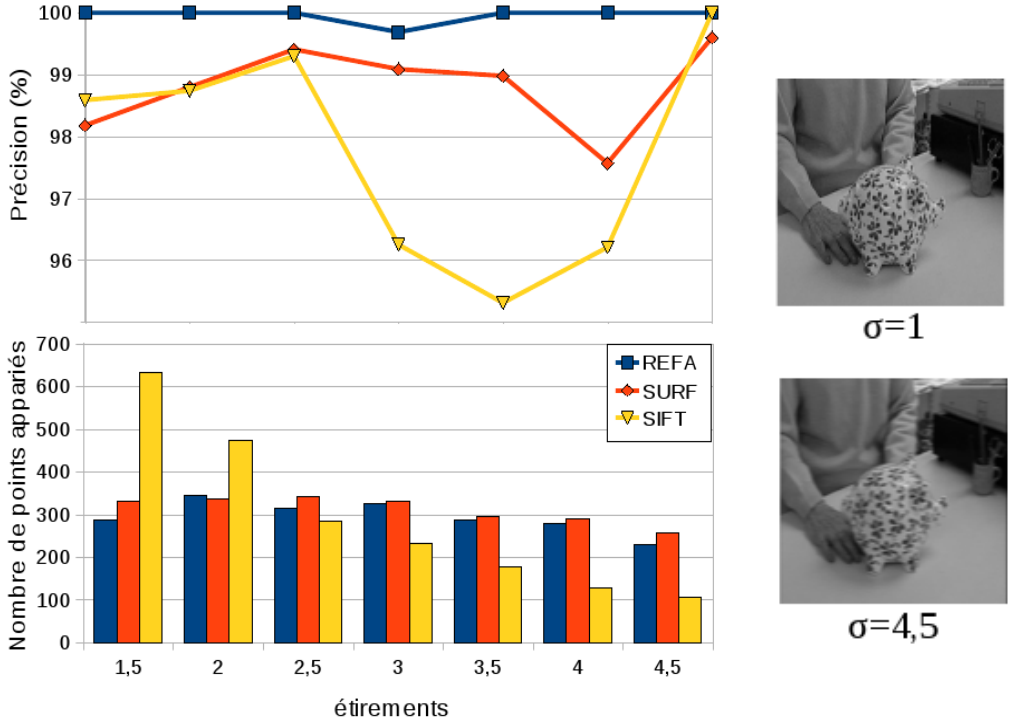


FIG. 3.18 – Précision et nombre de points appariés pour des étirements unidirectionnels sur les images Pig.

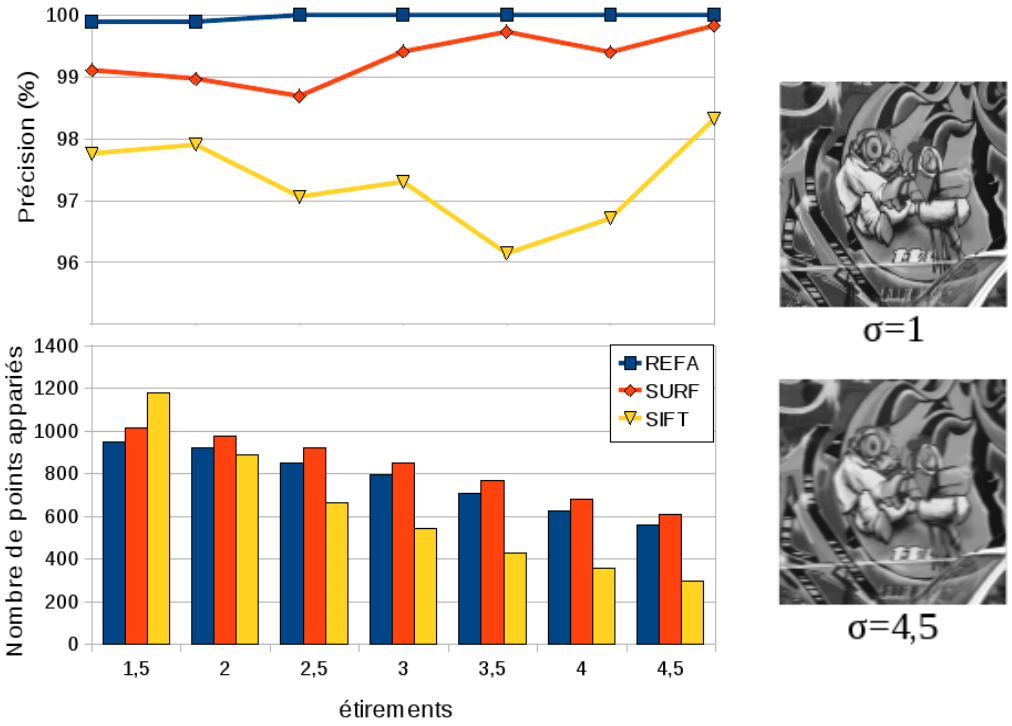


FIG. 3.19 – Précision et nombre de points appariés pour des étirements unidirectionnels sur les images Graffiti.

Notre méthode obtient les meilleurs résulats en termes de précision et d'invariance. La perte de quelques appariements en comparaison au SURF est compensé par une augmentation de la stabilité. Ces performances reposent en grande partie sur la mise à l'échelle de notre masque. En effet dans la cas d'un etiremment unidirectionnel l'information est étendue et notre descripteur s'adapte automatiquement.

– Les transformations de type **rotations** sont également étudiées. Les figures F.7 et F.8 regroupent les résultats obtenus pour les images Graffiti et Ubc. Par définition, la méthode SIFT est invariante à ce type de transformations.

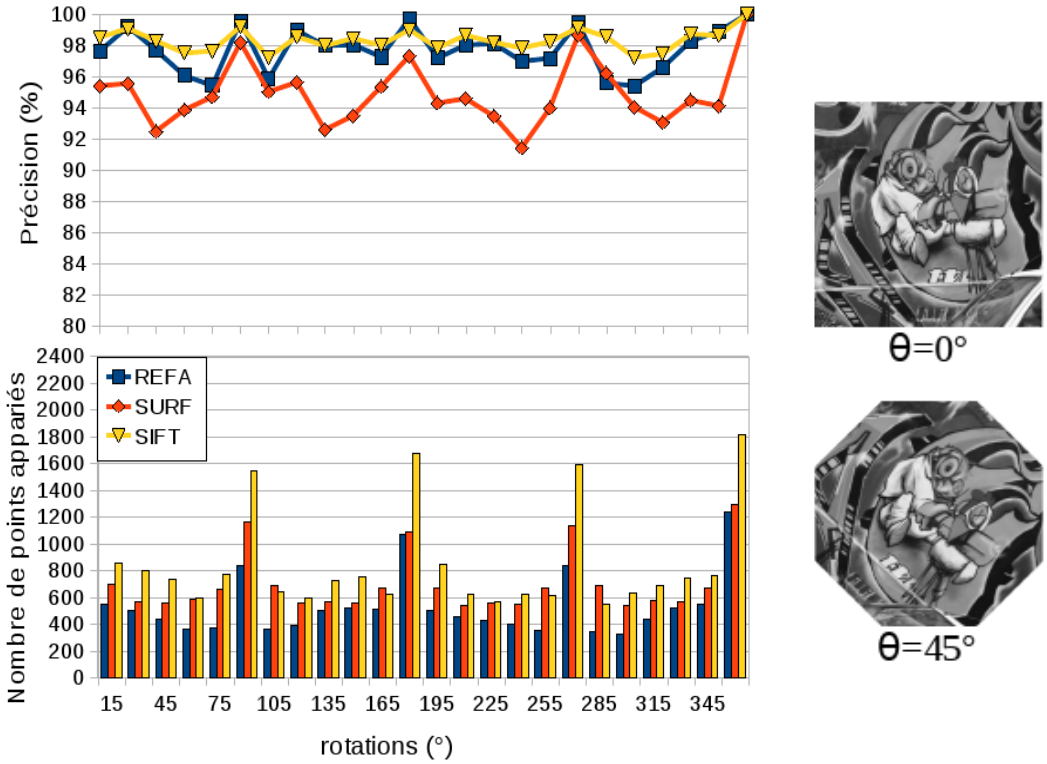


FIG. 3.20 – Précision et nombre de points appariés pour des transformations synthétiques de type rotation (images Graffiti).

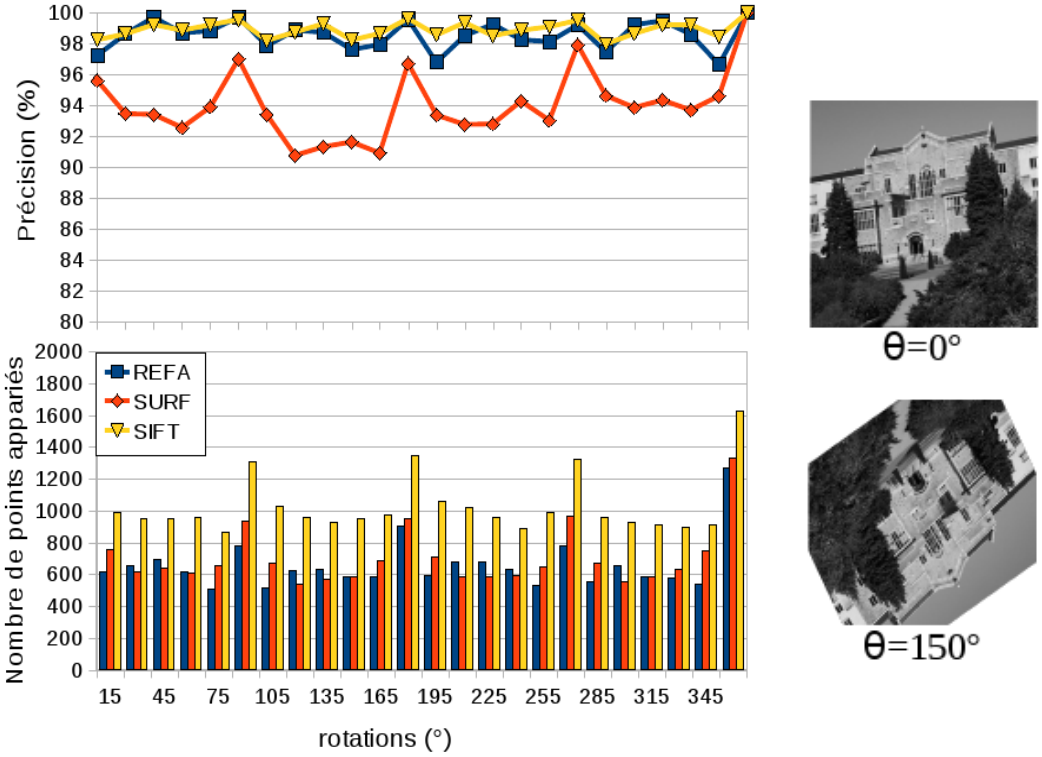


FIG. 3.21 – Précision et nombre de points appariés pour des transformations synthétiques de type rotation (images Ubc).

Nous observons que pour les images Ubc notre méthode présente une précision semblable à celle du SIFT et meilleure que SURF. D'autre part notre approche obtient des performances légèrement inférieures à celle du SIFT dans le cas d'images texturées (Graffiti) mais supérieures au SURF. La précision et la stabilité de notre méthode sont principalement dues au recalage de notre masque d'analyse ainsi qu'à sa forme. En effet nous avons détaillé au §3.2.1 les avantages d'une utilisation de masques circulaires (ou elliptiques) notamment dans le cas d'un recalage.

Nous proposons ensuite d'étudier les conséquences sur les différents critères d'observation, d'un couplage de plusieurs transformations telle que : une rotation de 45 ° et un changement d'échelle ou une rotation et un étirement.

– Nous analysons les transformations composées de **changements d'échelle et rotations**. Les figures F.9 et F.10 regroupent les résultats obtenus pour les images Graffiti et Beatles.

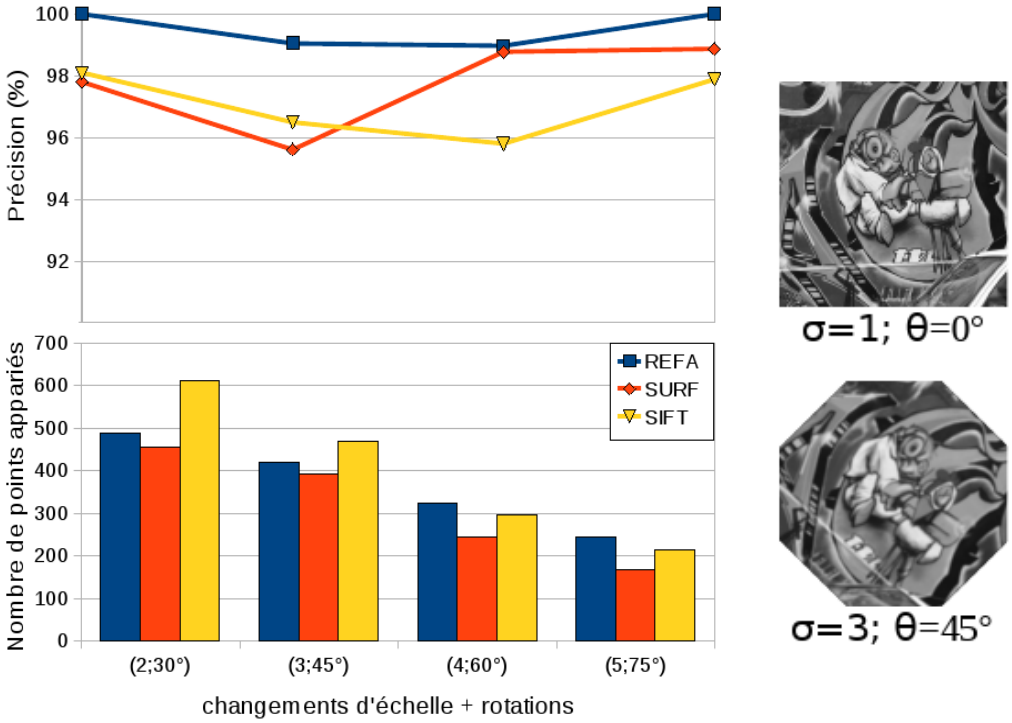


FIG. 3.22 – Précision et nombre de points appariés pour un couplage changements d'échelle/rotations (Graffiti).

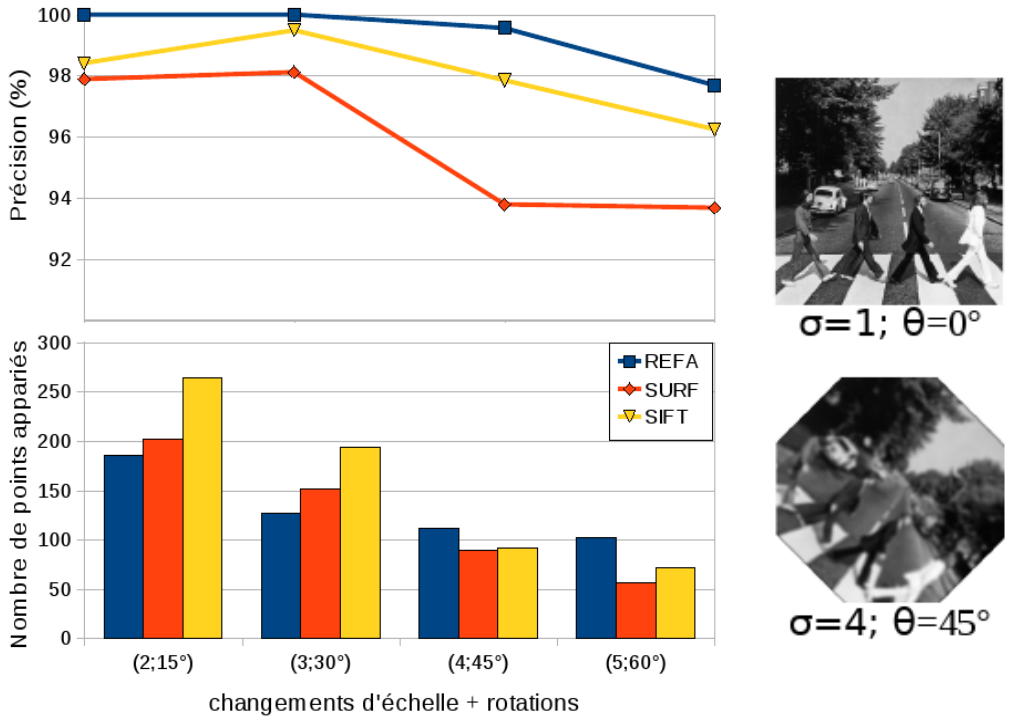


FIG. 3.23 – Précision et nombre de points appariés pour un couplages changements d'échelle/rotations (Beatles).

– Concernant le couplage de transformations de type **étirements et rotations**, les figures F.11 et F.12 présentent les résultats obtenus pour les images Lena et Pig.

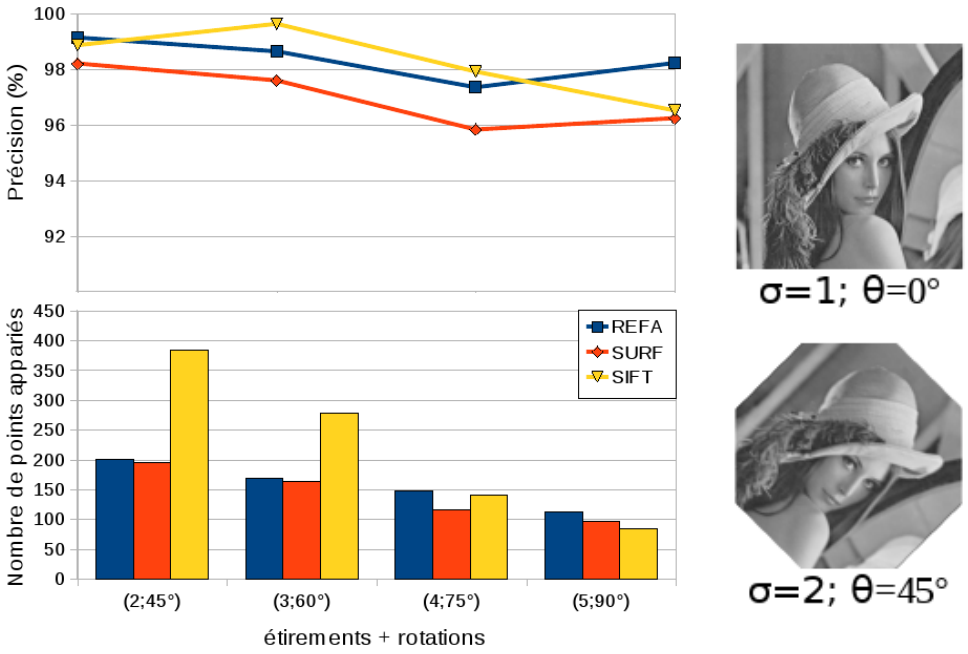


FIG. 3.24 – Précision et nombre de points appariés pour un couplage étirements et rotations (Lena).

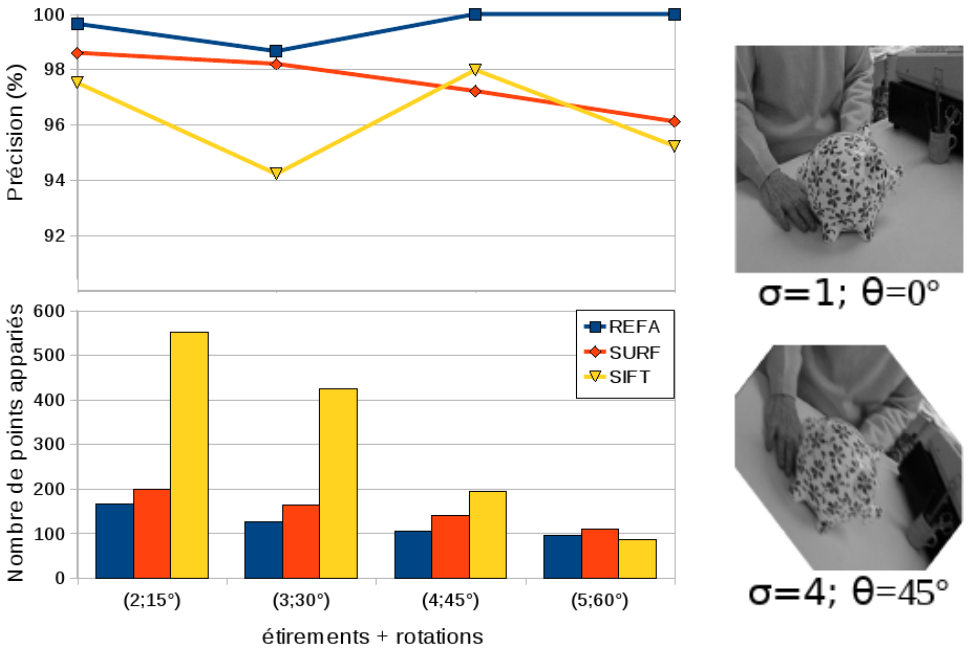


FIG. 3.25 – Précision et nombre de points appariés pour un couplages étirements et rotations (Pig).

Pour un couplage de transformations synthétiques, notre approche présente globalement les meilleurs résultats, aussi bien en terme de stabilité que de précision. Notre nombre de points appariés décroît également moins vite.

En conclusion, concernant les transformations synthétiques la méthode proposée possède une meilleure précision que SIFT et SURF ainsi qu'une meilleure robustesse aux types de transformations étudiées. Ces performances entraînent malheureusement une diminution du nombre de points appariés, faible vis à vis du SURF et plus importante vis à vis du SIFT. Néanmoins ce nombre décroît de façon moins importante pour notre approche.

3.5.3 Transformations réelles

Nous souhaitons également détailler les résultats de l'analyse de transformations réelles. Pour ce faire nous utilisons la base de donnée d'Oxford (voir Annexe A) regroupant l'ensemble des modifications pouvant subir une image. Nous proposons d'étudier la précision et le nombre d'appariement pour des transformations de type changements d'échelles et rotations, changements de point de vue, modifications de la compression, changements de luminosité ou bruitage de l'image.

– Concernant les transformations composées d'un couplage **changements d'échelles et rotations**, les courbes de la figure F.13 présentent les résultats obtenus pour les images Boat (voir Annexe A).

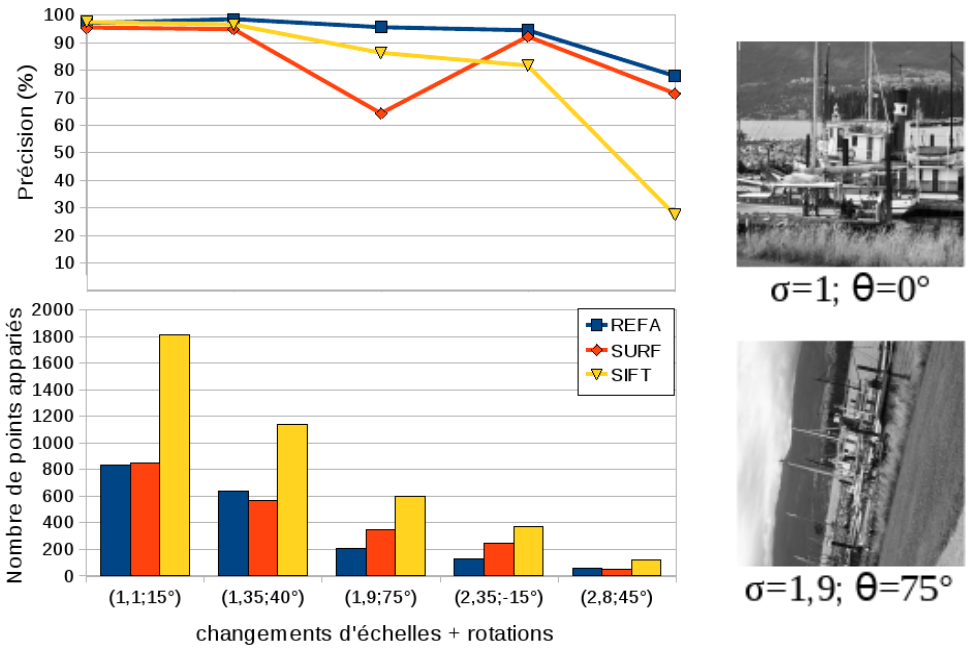


FIG. 3.26 – Précision et nombre de points appariés pour un couplage changements d'échelle et rotations (Boat).

Notre méthode présente une robustesse à ce type de transformations meilleure que celle de SIFT et SURF, ainsi qu'une très bonne précision. Ces résultats sont dus d'une part au recalage de notre masque d'analyse paliant aux rotations de l'image et d'autre part à la mise à l'échelle de notre descripteur. En effet notre région d'analyse dépend de l'échelle locale, notre masque suit donc la modification que peut subir cette dernière.

– Nous souhaitons également observer les conséquences de **modifications de compression JPEG** de l'image. Les courbes de la figure F.14 illustrent les résultats obtenus pour les images Ubc.

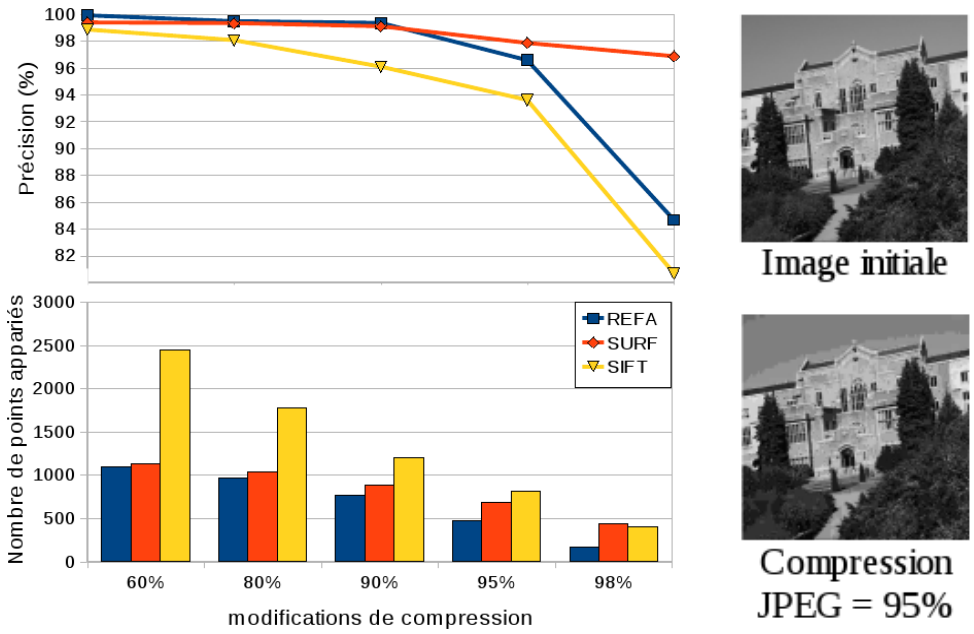


FIG. 3.27 – Précision et nombre de points appariés pour des modifications de compression de l'image (images Ubc).

Concernant les modifications de compression, une diminution des performances est à noter notamment pour des taux de 95% et 98%, pour lesquels SURF propose une meilleure robustesse. En effet, nous pouvons observer que notre approche est perturbée à partir d'un certain taux de compression des données. Notre descripteur se base sur un masque composé de dix-sept cercles, permettant ainsi d'extraire un maximum de détails dans le voisinage du point analysé. De ce fait une trop forte compression de l'information locale a pour conséquence la détérioration de la description et donc une diminution des résultats.

– Nous étudions des transformations de type **changements de luminosité**. Ces dernières sont obtenues par une variation de l'ouverture de la caméra, permettant notamment de réguler l'illumation du capteur. Les courbes de la figure F.15 présentent les résultats obtenus pour les images Leuven.

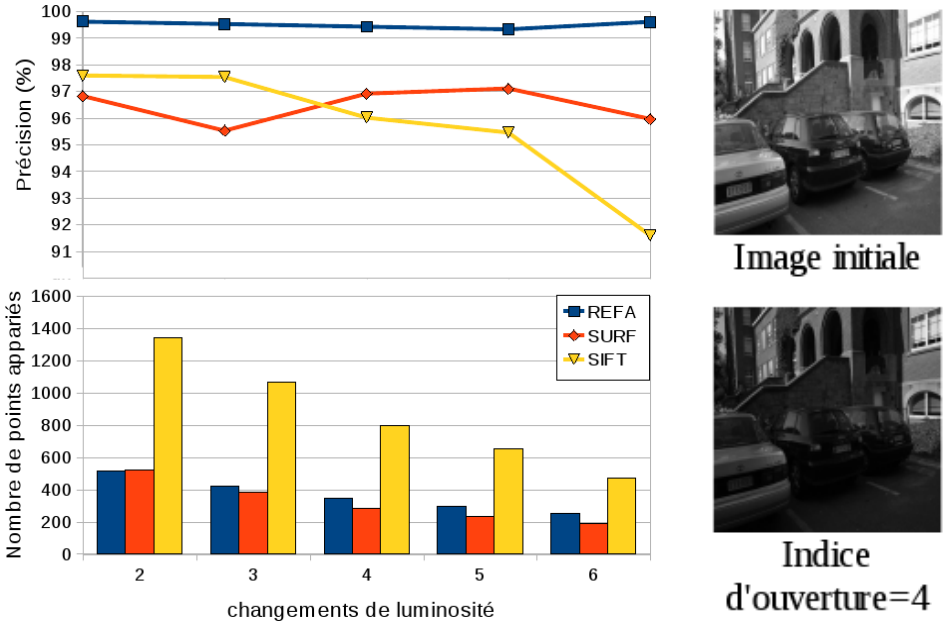


FIG. 3.28 – Précision et nombre de points appariés en fonction de l'indice d'ouverture de la caméra (plus l'indice est élevé, plus l'ouverture est petite).

– Le **bruitage** de l'image peut entraîner de fortes détériorations des résultats, nous proposons donc d'en observer les conséquences (figure F.16).

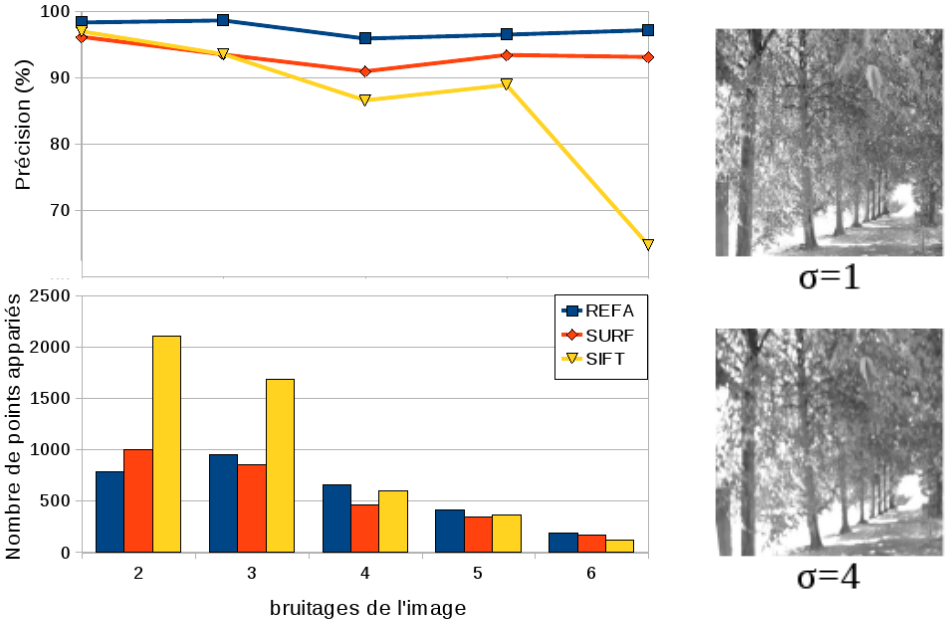


FIG. 3.29 – Précision et nombre de points appariés pour des ajouts de bruits gaussiens dans l'image (écart type $\sigma \in [\![2; 6]\!]$, images Trees).

Pour les transformations types changements de luminosités et bruitages de l'image, notre méthode obtient les meilleurs résultats en terme de précision. Le nombre de points appariés décroît moins rapidement que celui des deux autres approches. Cette robustesse est acquise d'une part grâce à la normalisation de notre descripteur (pour les changements de luminosité) et d'autre part grâce au seuillage des valeurs des gradients ainsi qu'au lissage global initial (pour le bruitage).

– Nous proposons d'analyser un dernier type de transformations : les **changements de point de vue**. Ce type de transformation est obtenu par rotation de la caméra suivant un plan normal à l'image. Les abscisses des figures F.17 et F.18, présentant les résultats obtenus pour les images Wall et Graffiti, correspondent donc aux angles utilisés pour ces rotations.

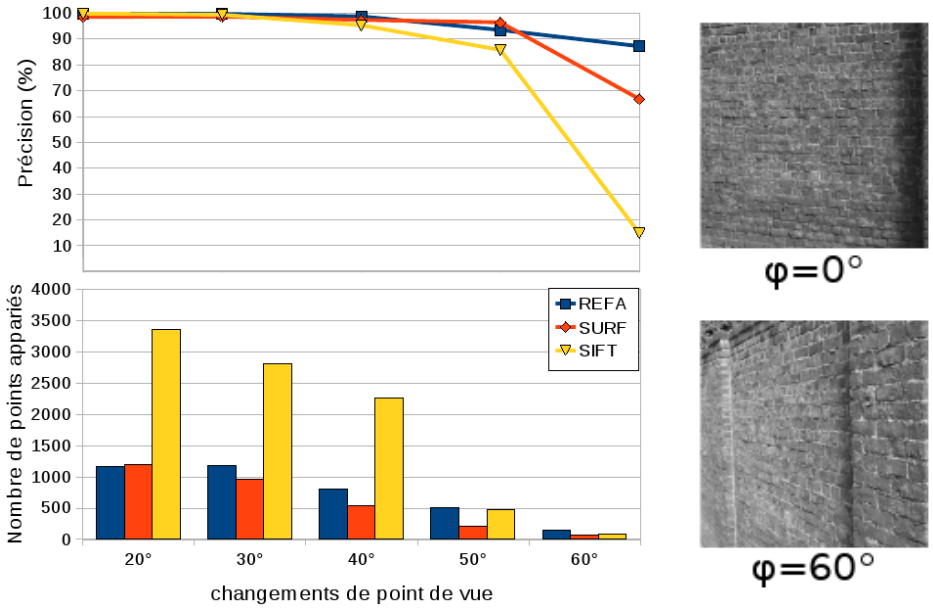


FIG. 3.30 – Précision et nombre de points appariés pour des changements de point de vue (images Wall).

Les tests sur les changements de point de vue montrent que notre approche possède la meilleure précision ainsi qu'un nombre de points appariés plus constant. Néanmoins ces résultats montrent également que de telles transformations restent difficiles à gérer notamment pour des images texturées de type Graffiti (figure F.18). En effet nous pouvons observer une forte diminution des performances pour des changements de point de vue dépassant 40 ° .

Concernant les transformations réelles, nous observons que les résultats obtenus avec notre approche confortent ceux établis au §3.5.2. En effet, notre méthode obtient une meilleure précision que le SIFT et le SURF pour des couplages rotations/changements d'échelle. Nous apportons également de nouvelles performances concernant les change-

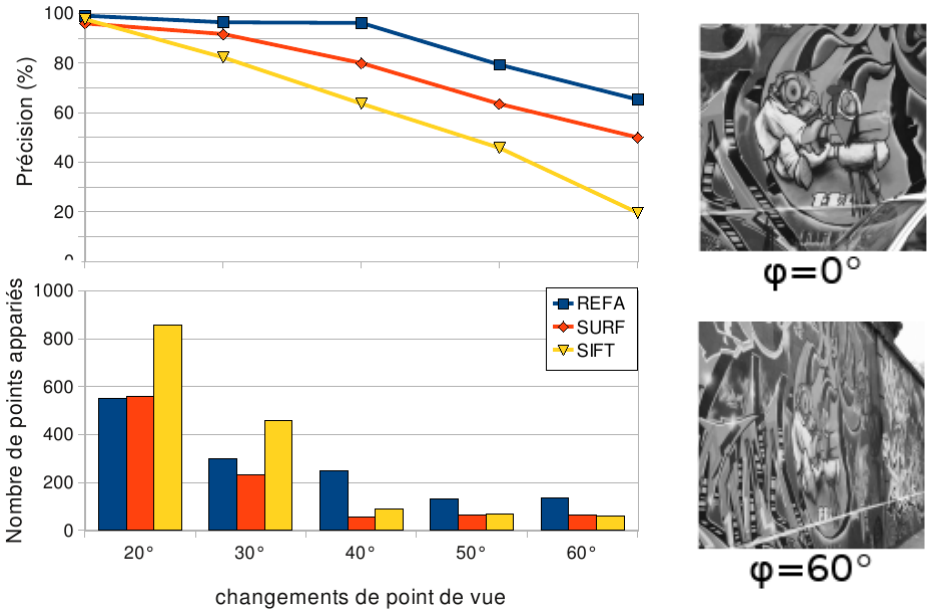


FIG. 3.31 – Précision et nombre de points appariés pour des changements de point de vue (images Graffiti).

ments de luminosité ainsi que le bruitage de l'image. Dans ces deux cas, notre descripteur présente la meilleure invariance et une précision d'appariement supérieure à celle des autres méthodes. Les modifications de compression de l'image peuvent néanmoins s'avérer perturbatrices pour notre méthode. Nous avons pu observer sur la courbe F.14 qu'il existe une limite en terme de taux de compression à partir de laquelle les performances de notre descripteur se dégradent fortement. Parmi les trois approches testées, seul le SURF propose une meilleure robustesse à ce type de modifications de l'image. Les tests effectués pour des changements de point de vue nous permettent de mettre en avant l'ensemble des outils et optimisations de notre méthode. Etant un type de transformations difficiles à gérer, les différents paramètres tels que le recalage, la mise à l'échelle et le seuillage des gradients prennent toute leur importance. Nous montrons par le biais des courbes F.17 et F.18 que notre approche permet d'obtenir la meilleure précision d'appariement ainsi qu'une certaine robustesse vis à vis de ce type de transformations. Nous pouvons également déduire de ces courbes une certaine invariance aux transformations affines. En effet en s'appuyant sur la figure 2.7 il apparait que le seul paramètre non géré par notre descripteur est la non-orthogonalité du repère. Néanmoins effectuant notre description d'un point de vue locale, il est possible de ramener les problèmes projectifs à des problèmes de similitudes et ainsi de gérer l'ensemble des transformations affines.

3.5.4 Influence de la détérioration des données

Afin détudier l'ensemble des performances de notre approche, nous observons les conséquences que peut avoir la détérioration des données. Pour ce faire le principe consiste à modifier notre seuil de validation α (équation 3.5) afin d'augmenter le nombre

de points pouvant être appariés, introduisant ainsi volontairement de 'faux candidats'. Les conséquences sont d'une part un accroîssement du nombre de possibilité de mise en correspondance, ce qui crée des faux appariements, et d'autre part une diminution de la précision. Notre étude se résume à trois types de transformation : un changement de luminosité, un couplage rotation/changement d'échelle et un changement de point de vue. Les figures 3.32, 3.33 et 3.34 regroupent les différents résultats comparatifs. Une analogie peut être faite entre ces courbes et des courbes ROC (*Receiver Operating Characteristic*) mettant en avant la sensibilité des classifieurs en fonction du seuil de discrimination. Dans notre analyse les classifieurs correspondent aux couples de points extraits et le seuil de discrimination s'applique sur le nombre de points pouvant être appariés.

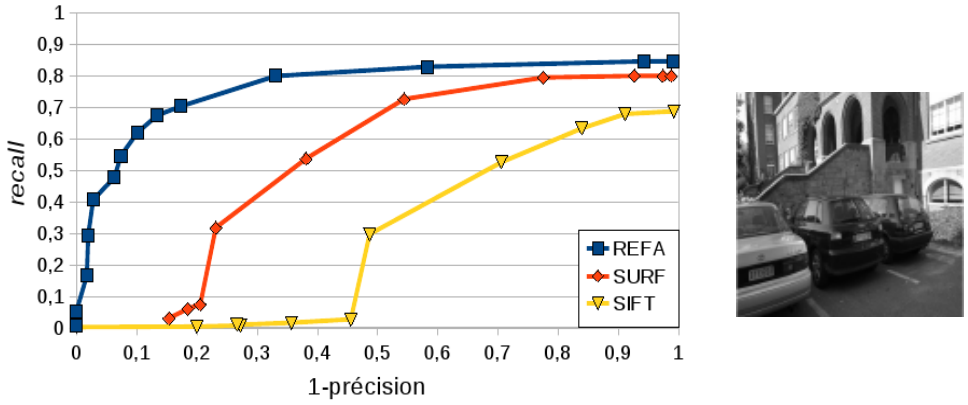


FIG. 3.32 – Taux d'appariements corrects dans l'étude de l'influence de la dégradation des données pour des changements de luminosité (Leuven).

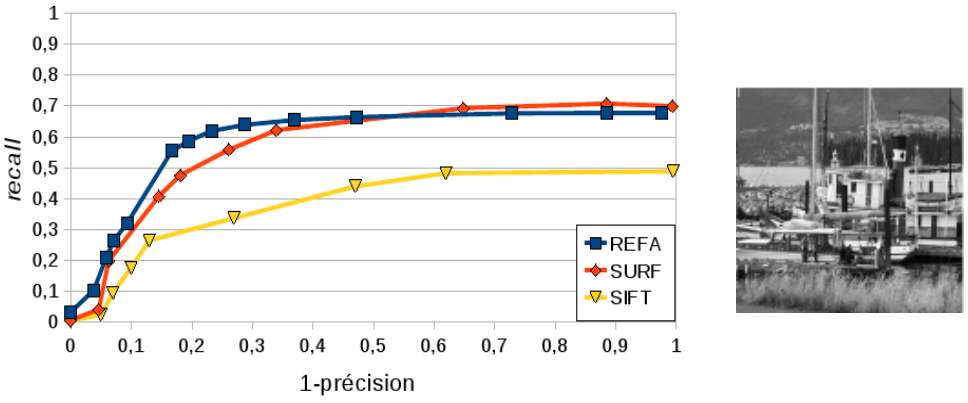


FIG. 3.33 – Taux d'appariements corrects dans l'étude de l'influence de la dégradation des données pour des rotations et changements d'échelle (Boat).

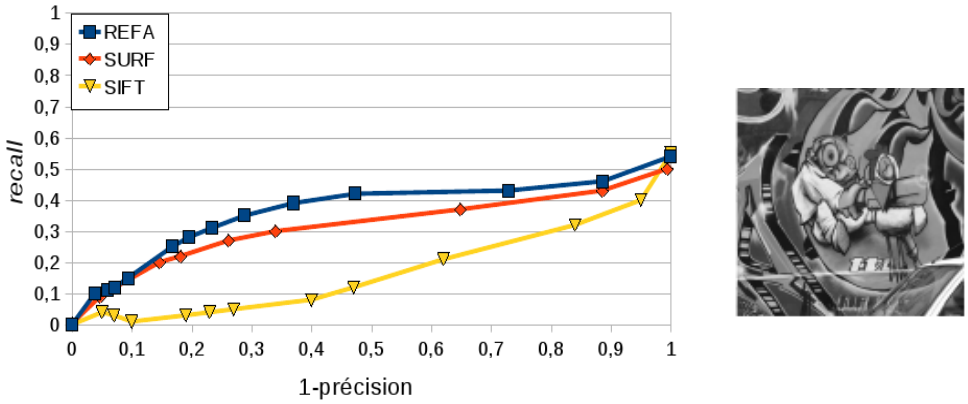


FIG. 3.34 – Taux d'appariements corrects dans l'étude de l'influence de la dégradation des données pour des changements de point de vue (Graffiti).

Notre approche présente une meilleure robustesse vis à vis de la détérioration des données. L'augmentation du critère 1-précision (ajout de faux candidats) perturbe l'ensemble des méthodes comparées. Néanmoins l'approche REFA obtient des performances supérieures à celles du SIFT et du SURF en terme de taux d'appariement correct (*recall*). L'ensemble des procédés (recalage, mise à l'échelle, masque à base elliptique...) utilisés pour la construction de notre descripteur permet donc de créer une caractérisation du voisinage plus pertinente.

3.5.5 Synthèse des résultats obtenus

Afin de visualiser plus aisément les différents résultats énoncés au cours de ce chapitre, nous les regroupons et en donnons une synthèse. Cette dernière se base d'une part sur l'ensemble des tests présentés précédemment et d'autre part sur divers résultats non énoncés mais nous permettant d'étoffer et de confirmer les performances de notre approche. Les regroupements se basent sur le type de transformations étudiées dont les noms sont abrégés de la façon suivante : 'BO' pour la Base d'Oxford ('w' : wall, 'g' : graffiti, 'l' : leuven, 'b' :boat, 'u' : ubc, 't' :trees) et 'BS' pour la Base constituée des transformations Synthétiques ('e' : changements d'échelles, 'et' : étirements, 'r' : rotations, 're' : couplage rotations/changements d'échelles, 'ret' : couplage rotations/étirements). La figure 3.35 présente à la fois la précision et le taux de mise en correspondance obtenus. Ce dernier critère d'observation s'appuie sur le nombre de points appariés et permet de le quantifier par rapport au nombre de points pouvant être appariés. En effet il paraît plus logique d'observer ce ratio du fait qu'une approche telle que SIFT, ayant un très grand nombre de points détectés (notamment pour les premières transformations), présente un nombre d'appariement forcément plus élevé que notre approche ou celle du SURF.

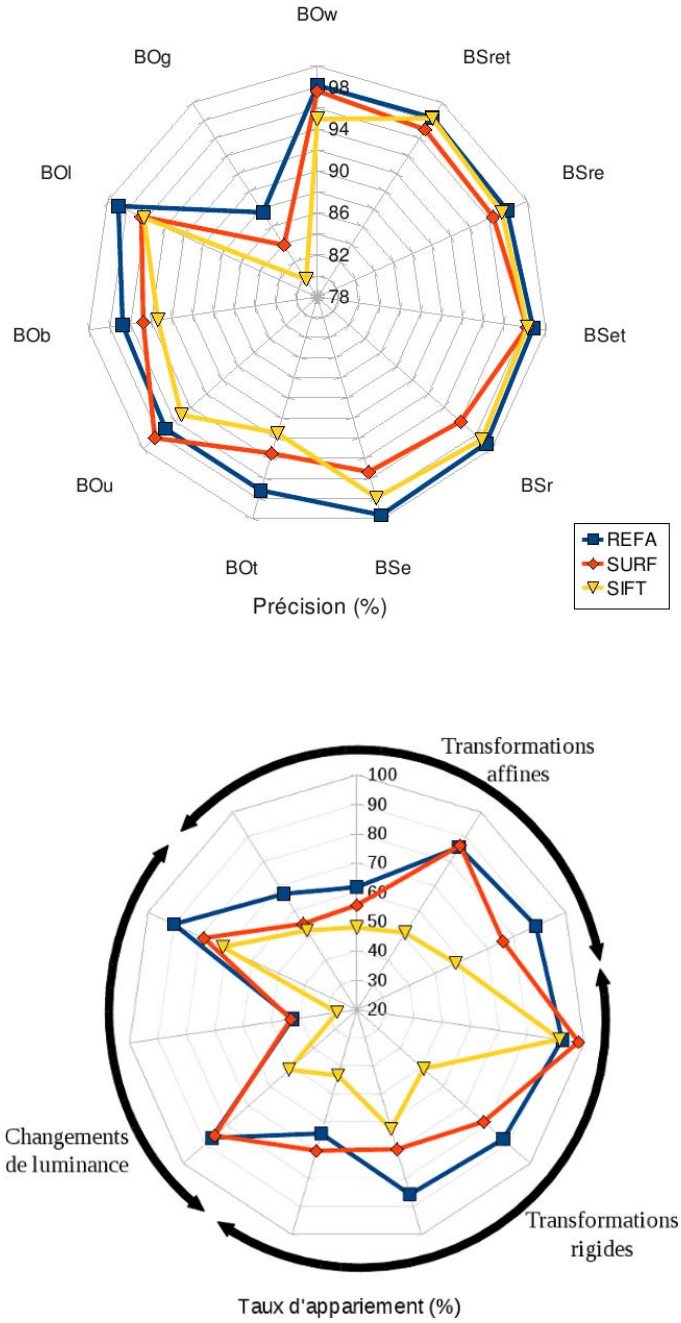


FIG. 3.35 – Synthèse des résultats obtenus en terme de précision (haut) et de taux d'appariement (bas) définis respectivement en 3.6 et 3.7.

Cette synthèse nous permet de valider notre approche, obtenant généralement des résultats supérieurs au SIFT et au SURF. En détaillant le premier graphique nous observons que notre méthode présente la meilleure précision pour la quasi totalité des transformations, seule la modification de la compression (Base Oxford ubc) perturbe plus notre approche que celle du SURF. Concernant les transformations comportant

des rotations nous obtenons des résultats similaires à SIFT, répondant à l'objectif fixé au §3.5.2 (partie sur les rotations). Nous pouvons également remarquer une certaine stabilité dans nos performances quelque soit le type de tranformation étudiée. En effet notre précision reste comprise entre $96,76\%$ et $99,66\%$ représentant une variation de $2,2\%$ en terme de performance, contre $4,99\%$ pour SURF et 8.76% pour SIFT. D'un point de vue taux d'appariement le constat émis précédemment reste le même. Notre approche présente les meilleurs résultats pour la majorité des transformations étudiées, seuls les rotations et le couplage rotations/changements d'échelle entraînent des performances inférieures. Nous pouvons donc affirmer que notre méthode, à nombre de points détectés égal, propose le plus grand nombre de points appariés pour la majorité des tests.

En définitive au vu des nombreux résultats proposés ci-dessus, il apparaît que notre approche présente le meilleur taux d'appariement pour la quasi totalité des tests ainsi qu'une très bonne précision. En s'appuyant sur sa comparaison avec le SIFT et le SURF, nous validons donc notre méthode tout en illustrant sa stabilité et ses performances.

3.5.6 Résultats complémentaires

Nous avons validé notre approche par un grand nombre d'études comparatives. Néanmoins, certains compléments peuvent être apportés afin d'analyser pleinement le gain en terme de performance qu'apporte notre méthode. Pour ce faire, nous comparons dans un premier temps, les trois méthodes (SIFT, SURF et REFA) en limitant le nombre de points détectés. Nous ne conservons que les k points présentant le meilleur score de détection (avec k le nombre minimum de points détectés par une des trois méthode). Afin d'éviter l'accumulation de courbes, nous nous limitons aux figures 3.36 et 3.37, présentant les résultats obtenus pour des changements de point de vue. Les autres résultats sont listés en annexe F.

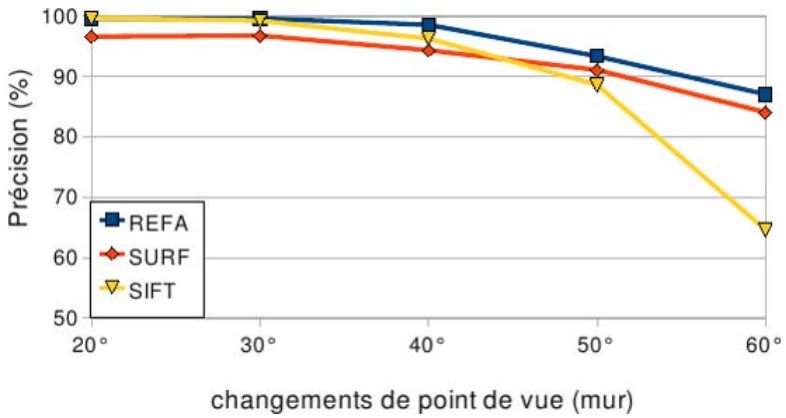


Fig. 3.36 – Précision pour des changements de point de vue avec seuillage du nombre de points détectés (images Wall).

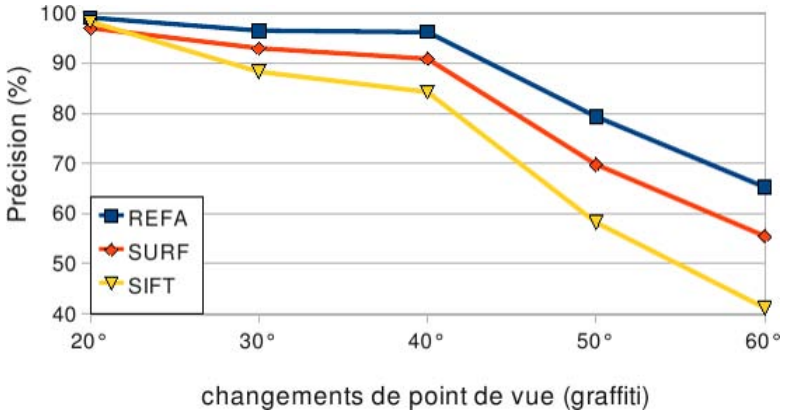


FIG. 3.37 – Précision pour des changements de point de vue avec seuillage du nombre de points détectés (images Graffiti).

Dans un second temps, nous fournissons aux trois approches, une liste de points initiaux identique, afin de mettre en avant les avantages qu'apportent notre description elliptique adaptative. Les figures 3.38 et 3.39 présentent les précisions des méthodes, pour des tranformations de type changements de point de vue.

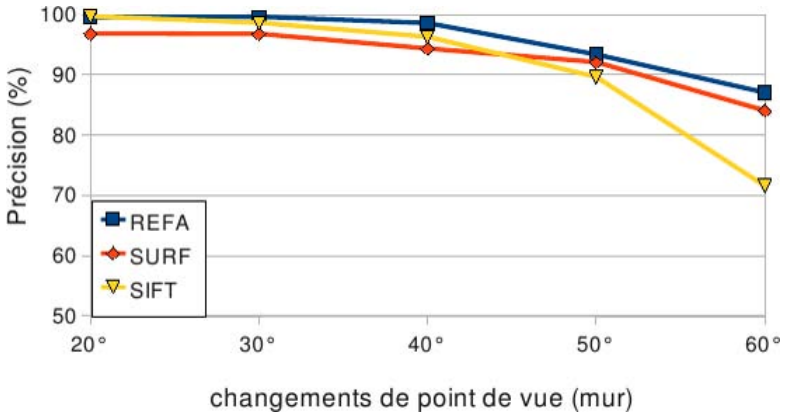


FIG. 3.38 – Précision pour des changements de point de vue avec des points initiaux identiques (images Wall).

A travers ces deux études comparatives, il apparaît que l'ajout d'une analyse elliptique permet de mieux observer l'information locale du point d'intérêt. D'autre part l'aspect adaptatif de notre descripteur lui octroie un meilleur suivi de la déformation de l'information. Notre approche apporte donc un gain non négligeable en terme de précision et de taux d'appariements.

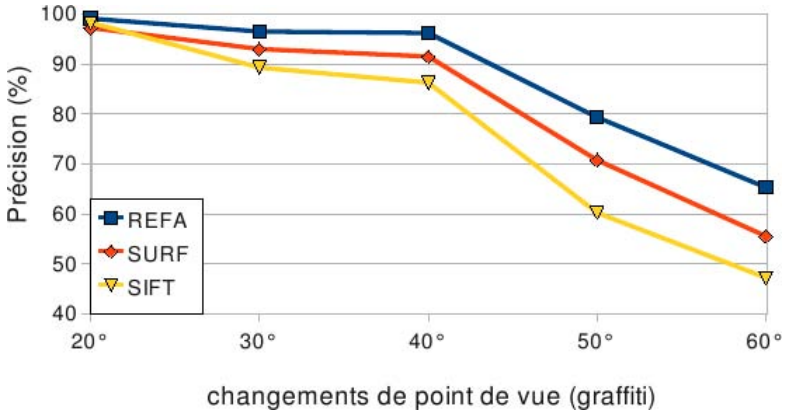


FIG. 3.39 – Précision pour des changements de point de vue avec des points initiaux identiques (images Graffiti).

Les deux méthodes utilisées pour notre étude s'appuient sur un description carré du voisinage. Afin de finaliser la validation de notre approceh, nous proposons d'ajouter le descripteur DAISY à nos observations. En effet, son exploration circulaire se rapproche de notre analyse elliptique. Il s'avère donc judicieux de les comparer. Les figures 3.40 et 3.41 illustrent les résultats ainsi obtenus.

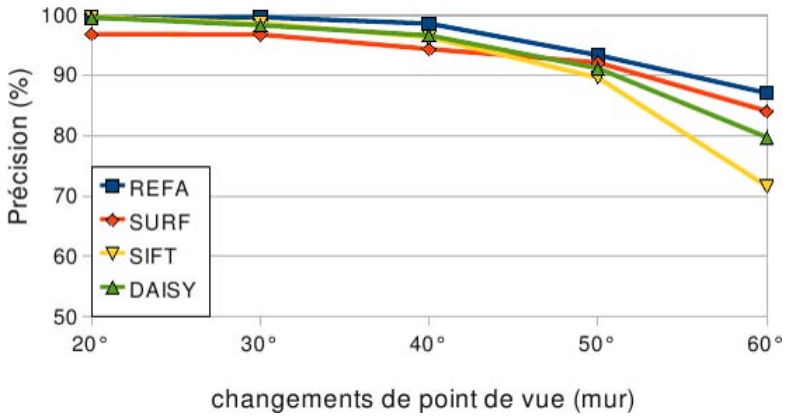


FIG. 3.40 – Précision pour des changements de point de vue avec l'insertion de la méthode DAISY (images Wall).

Nous observons une amélioration de la précision lors de l'utilisation d'un masque circulaire. Néanmoins, l'approche DAISY reste moins performante que notre méthode pour ce type de transformations. En définitive, notre analyse elliptique adaptative permet de mieux suivre les modifications, notamment non-uniforme, de l'information locale et par conséquent de proposer une description plus pertinente et plus précise du voisinage du point d'intérêt.

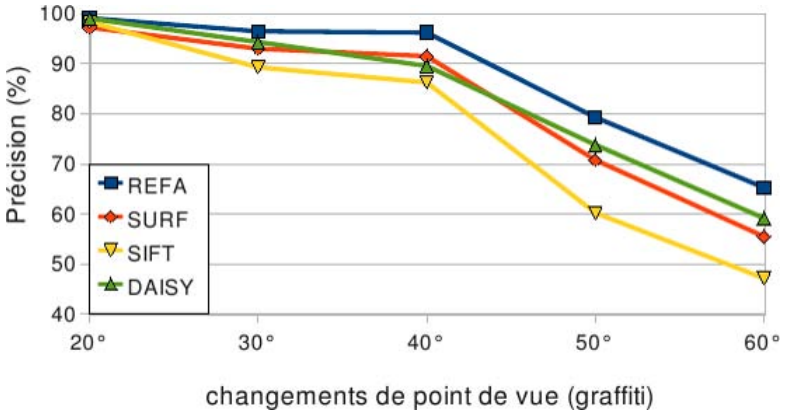


FIG. 3.41 – Précision pour des changements de point de vue avec l'insertion de la méthode DAISY (images Graffiti).

3.5.7 Perspectives d'utilisation de la méthode REFA

Les méthodes SIFT et SURF sont couramment utilisées dans des applications de type reconstruction 3D, stéréo-vision ou encore suivi d'objets par exemple. Elles permettent de fournir des appariements de points dont une des utilités est d'estimer l'homographie existante entre deux images. Nous avons observé au §3.5.5 que notre approche présente de bons résultats concernant la précision et la stabilité. Nous proposons donc de comparer les estimations des homographies basées sur SIFT, SURF et REFA. La mise en place de ces estimations s'appuie sur l'utilisation de la méthode RANSAC détaillée en annexe E. Les tableaux 3.1 et 3.2 regroupent les résultats obtenus pour différentes transformations. Nous analysons d'une part, le pourcentage de points validant l'estimation de l'homographie et d'autre part, le taux d'erreur (basé sur la distance euclidienne en pixel) obtenu lors de la reprojection des points de la seconde image dans l'image initiale.

	BOg ($\varphi = 30^{\circ}$)		BOb ($\sigma = 1,35$ et $\theta = 40^{\circ}$)	
	erreur de reprojection (%)	points validant l'estimation (%)	erreur de reprojection (%)	points validant l'estimation (%)
REFA	0,06	96,72	0,54	93,23
SURF	3,46	96,34	1,69	86,03
SIFT	2,98	92,35	2,06	88,24

TAB. 3.1 – Résultats de l'estimation des matrices d'homographie pour des transformations de type changements de points de vue et couplage rotations/changements d'échelle.

	BOt ($\sigma = 4$)		BOl (indice d'ouverture = 4)	
	erreur de reprojection (%)	points validant l'estimation (%)	erreur de reprojection (%)	points validant l'estimation (%)
REFA	0,61	97,13	0,47	98,21
SURF	0,66	92,1	0,69	94,53
SIFT	0,58	96,56	0,95	90,84

TAB. 3.2 – Résultats de l'estimation des matrices d'homographie pour des transformations de type bruitages de l'image et changements de luminosité.

L'analyse du taux d'erreur d'estimation et du pourcentage de points pris en considération lors de la validation de l'homographie permet de valider qualitativement les couples de points retournés par chaque méthode. Pour les deux premières transformations étudiées, notre approche obtient un taux d'erreur inférieur à ceux du SIFT et du SURF, ainsi qu'un nombre de points supérieur. Concernant les bruitages de l'image, SIFT présente un taux d'erreur plus faible, néanmoins nous compensons cette différence par un pourcentage de points supérieur. Pour la dernière transformation étudiée, notre méthode propose les meilleurs résultats.

Nous avons démontré que notre méthode REFA présente de bons résultats lors de la mise en correspondance de deux images. Dans l'optique d'étudier pleinement le potentiel de notre approche et d'élargir son domaine d'utilisation, nous proposons d'analyser son comportement en présence de séquences vidéo. Les tests mis en place consistent à recaler la 50ème image dans sa séquence initiale. Pour ce faire nous nous appuyons sur deux séquences vidéo, chacune composée de 100 images issues de la plateforme PAVIN (séquences 1 et 2 détaillées en annexe B). Le principe est d'apparier la 50ème image avec chaque image de la séquence, puis d'observer au fur et à mesure le taux d'appariement Ta_{50} (nombre d'appariements identiques à ceux de la 50ème image divisé par le nombre de correspondance de la 50ème image). Les figures 3.42 et 3.43 illustrent les resultats obtenus pour le recalage de la 50ème image des séquences 1 et 2.

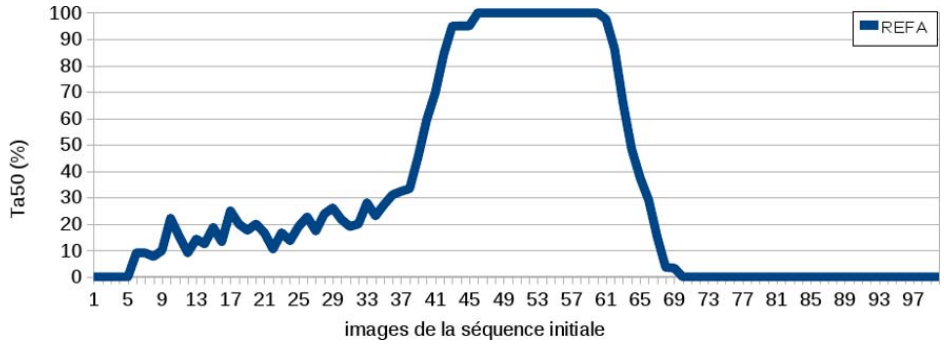


FIG. 3.42 – Précision obtenue lors du recalage de la 50ème image dans sa séquence initiale (séquence 1 de PAVIN).

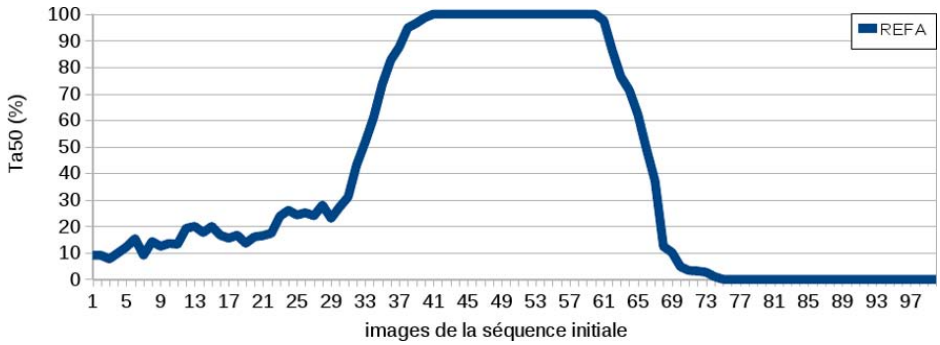


FIG. 3.43 – Précision obtenue lors du recalage de la 50ème image dans sa séquence initiale (séquence 2 de PAVIN).

La thématique des véhicules intelligents intègre un certain nombre de procédés et notamment la localisation dans le temps. Cette dernière s'appuie essentiellement sur le recalage des images, observées par le véhicule lors de son déplacement, dans une séquence initialement apprise. Pour un tel processus, les performances présentées en figures 3.42 et 3.43 ne donnent pas satisfaction. En effet, nous n'obtenons pas 1 maximum de précision (cas idéal), mais 14 pour la séquence 1 et 20 pour la séquence 2. Par conséquent il nous est impossible de recaler avec précision l'image I dans le temps. Nous démontrons donc qu'une méthode se limitant à une analyse spatiale ne permet pas d'obtenir les résultats souhaités. La principale raison réside dans le fait qu'un point d'intérêt spatial, possèdant un déplacement dans le temps, apparaît dans une succession d'images mais son voisinage reste quasi inchangé. Notre processus de description local ne prenant en compte que l'information spatiale, les histogrammes de gradients orientés calculés dans les images t,$t+1$,...,$t+n$ (où n correspond à la dernière image contenant le point analysé) restent semblables. Au final, cette analyse fournissant n fois le même descripteur, les mises en correspondance de l'image I avec des images très proches (de la 40ème à la 60ème dans ces tests) procurent donc des résultats quasi identiques.

Dans le but de remédier à ce problème nous détallions au chapitre suivant une généralisation de la méthode REFA. Les différentes modifications dues à l'ajout des données temporelles, notamment vis à vis de notre masque d'analyse et de la construction de nos histogrammes, seront détaillées puis validées. Nous étudierons un certain nombre de tests basés sur des séquences issues, d'une part, de la plateforme PAVIN (Annexe B), et d'autre part, de simulateurs (Annexe C et D).

4 Méthode REFA3D : analyse robuste 2D+t de séquences vidéo

L'information temporelle permet notamment d'analyser les déplacements des points dans une séquence vidéo. Il nous faut par conséquent étendre notre approche au domaine spatio-temporel, tout en conservant les différentes contraintes que nous nous sommes fixées au chapitre 3 (robustesse aux transformations de l'image, taux d'appariement et précision). Nous allons dans un premier temps étudier les méthodes existantes pour ce type de problématique, puis nous proposerons une méthode conservant dans la mesure du possible les différents paramètres détaillés au chapitre précédent. Nous terminerons cette partie par des résultats comparatifs entre notre approche REFA3D, le SIFT3D et le couplage HOG/HOF.

4.1 Méthodes existantes

4.1.1 Détecteurs utilisés pour une analyse spatio-temporelle

De nombreuses approches ont vu le jour depuis 2003, proposant une extraction des points d'intérêt spatio-temporels. Ces différentes méthodes sont présentées, dont celle de Laptev et Lindberg en 2003 [64], Dollar et al. en 2005 [38] et Willems et al. en 2008 [111]. Il est également possible d'utiliser une approche flot optique [4],[10],[51] pour extraire les primitives spatio-temporelles.

4.1.1.1 Détecteur proposant une extension spatio-temporelle

Introduit en 2003 par Laptev et Lindberg [64], le Harris3D se base sur les équations 2.17 et 2.19. Les auteurs introduisent l'analyse temporelle dans la matrice de Harris afin d'obtenir le tenseur de structure suivant :

$$\mathbf{M} = g_{\sigma,\tau} * \begin{bmatrix} I_x^2 & I_x I_y & I_x I_t \\ I_x I_y & I_y^2 & I_y I_t \\ I_x I_t & I_y I_t & I_t^2 \end{bmatrix}, \tag{4.1}$$

avec $g_{\sigma,\tau}$ la fonction gaussienne spatio-temporelle, définie par un écart-type spatial σ et par un écart-type temporel τ. Le critère de Harris devient par conséquent égal à :

$$k_H(\mathbf{x}) = \det(\mathbf{M}) - \alpha \mathrm{trace}(\mathbf{M})^3 \tag{4.2}$$

Les auteurs démontrent que les résultats sont optimaux pour $(\sigma, \tau) \in [4; 64] \times [2; 4]$.

En 2005, Dollar et al. [38] proposent d'ajouter un filtrage temporel à un détecteur classique de type Harris. En décomposant l'équation 2.17 et en la couplant avec des filtres de Gabor, la fonction de réponse du détecteur devient :

$$R = (g_\sigma * h_{ev} * I(x, y, t))^2 + (g_\sigma * h_{od} * I(x, y, t))^2, \tag{4.3}$$

avec $h_{ev}(t; \tau) = -cos(8\pi t)e^{-t^2/\tau^2}$ et $h_{od}(t; \tau) = -sin(8\pi t)e^{-t^2/\tau^2}$ les réponses impulsionnelles des filtres temporels, et g_σ celle du filtre spatial. Une illustration de ces deux filtres est présentée dans la figure 4.1, permettant de visualiser l'analyse temporelle ainsi créée.

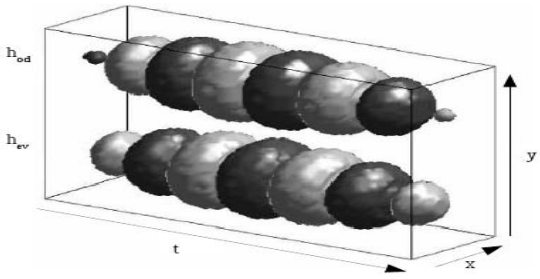


FIG. 4.1 – Représentation des deux filtres h_{ev} et h_{od}.(image extraite de [38])

Willens et al. [111] reprennent en 2008 l'idée générale de Laptev et Lindberg afin de l'appliquer à la matrice Hessienne (équation 2.1). En injectant la composante temporelle dans l'équation 2.45, cette dernière devient :

$$\mathbf{H}(\mathbf{x}; \sigma, \tau) = \begin{bmatrix} L_{xx}(\mathbf{x}; \sigma, \tau) & L_{xy}(\mathbf{x}; \sigma, \tau) & L_{xt}(\mathbf{x}; \sigma, \tau) \\ L_{xy}(\mathbf{x}; \sigma, \tau) & L_{yy}(\mathbf{x}; \sigma, \tau) & L_{yt}(\mathbf{x}; \sigma, \tau) \\ L_{xt}(\mathbf{x}; \sigma, \tau) & L_{yt}(\mathbf{x}; \sigma, \tau) & L_{tt}(\mathbf{x}; \sigma, \tau) \end{bmatrix} \tag{4.4}$$

Les points d'intérêt sont extraits par maximisation du déterminant de cette matrice. Une modification concernant l'approximation faite par Bay et al. [13] pour les masques de convolution (figure 2.18) a été apportée afin d'intégrer la notion temporelle. La figure 4.2 présente les nouvelles approximations des filtres gaussiens.

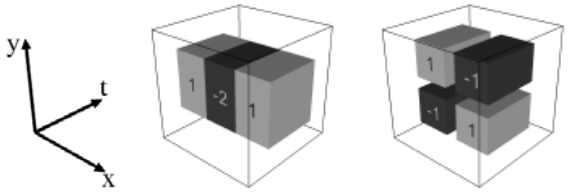


FIG. 4.2 – Représentation des filtres gaussiens dérivatifs spatio-temporels unidirectionnels (à gauche) et bi-directionnels (à droite).(image extraite de [111])

De plus en exportant les équations 2.34 et 2.35 au domaine spatio-temporel, les échelles caractéristiques sont déterminées par :

$$(\sigma_c, \tau_c) = \underset{\sigma,\tau}{\operatorname{argmax}}(\sigma^{2p}\tau^{2q}L_{xx}L_{yy}L_{tt}) \tag{4.5}$$

En s'appuyant sur les résultats obtenus, les auteurs préconisent une valeur de p égale à $\frac{5}{2}$ et une valeur de q égale à $\frac{5}{4}$. L'image 4.3 illustre cette détection ainsi que le voisinage du point d'intérêt.

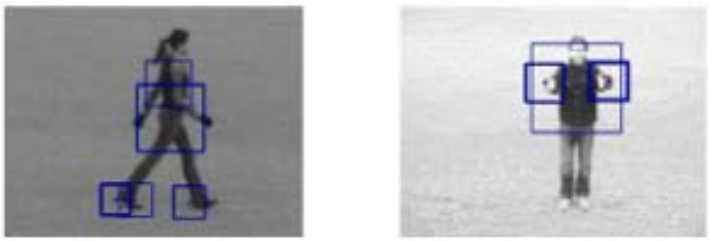


FIG. 4.3 – Exemple de détections spatio-temporelles (image extraite de [111]).

4.1.1.2 Couplage détecteur classique et flot optique

Introduiteen 1981 par Horn et Schunck [51], puis repris notamment par Adelson et Bergen en 1985 [4], le flot optique permet d'extraire des variations dans une séquence d'images en terme de déplacement de points. Une analyse exhaustive des différentes méthodes existantes a été proposée par Barron et al. en 1994 [10]. Ces différentes approches s'appuient sur l'hypothèse que la luminance d'un objet est constante d'une image à l'autre. D'un point de vue générale, cela revient à chercher en un point $\mathbf{x}$ de l'image I_t, le vecteur déplacement $\mathbf{v} = (v_x; v_y)$ tel que, à l'image I_{t+1}, $\mathbf{x} + \mathbf{v}$ permet d'obtenir la même valeur de niveau de gris.

Afin de déterminer ce vecteur, l'auteur s'appuie sur le fait que le déplacement d'un point entre deux images successives est faible et le représente par :

$$I(x, y, t) = I(x + v_x + \Delta t, y + v_y + \Delta t, t + \Delta t), \tag{4.6}$$

où Δt correspond au déplacement temporel. En appliquant les séries de Taylor à l'image ayant subit un faible déplacement, il est possible après identification d'obtenir l'équation suivante :

$$\frac{\partial I}{\partial x}v_x + \frac{\partial I}{\partial y}v_y + \frac{\partial I}{\partial t} = 0, \tag{4.7}$$

et présente deux inconnues, il est donc nécessaire d'utiliser l'hypothèse que des points proches possèdent des mouvements similaires. Cette dernière permet d'obtenir d'avantage d'équations, il est donc possible d'en extraire les composantes $\mathbf{v}_x$ et $\mathbf{v}_y$ du vecteur flot optique $\mathbf{v}$.

Afin d'accroître les performances de ce type de détecteur, une analyse multi-échelles est envisagée. D'un point de vue pratique, cela revient à faire tout d'abord des mesures sans tenir compte des détails fins de l'image, puis d'affiner progressivement l'analyse à des échelles plus locales. Le figure 4.4 illustre l'analyse par flot optique d'une scène et permet d'observer, d'un point de vue local, les vecteurs de déplacements.

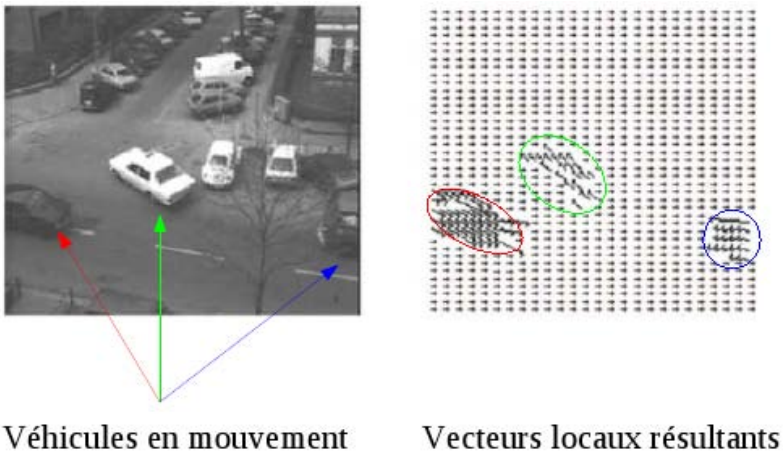


FIG. 4.4 – Représentation d'une scène par flot optique. Chaque véhicule en mouvement est représenté par un vecteur local.

L'extraction des primitives se limite donc aux zones définies par les vecteurs de déplacements et s'appuie sur des méthodes classiques telles que Harris, DoG ou autres. Ce couplage permet d'extraire uniquement des points d'intérêt subissant une modification temporelle. Ke et al. [59] proposent notamment en 2005 une approche se basant sur un couplage flot optique/ondelettes de Haar, appliqué sur des images intégrales.

4.1.2 Descripteur s'appuyant sur une analyse 2D+t

La description spatio-temporelle se base généralement sur le couplage ou l'extension 2D+T de méthodes existantes telles que SIFT ou SURF. Nous étudions ces descripteurs en s'inspirant du référencement publié par Wang et al. en 2009 [110]. Nous détaillerons le HOG3D introduit par Klaser et al. en 2008 [61] et la méthode N-SIFT de Cheung et Hamarneh [23] correspondant tout deux à une extension du SIFT. Nous analyserons également le couplage HOG/HOF (histogrammes de flot optique) décrit par Laptev et al. en 2006-2007 [65][63] ainsi que l'extension du SURF proposée en 2008 par Willems et al. [111].

4.1.2.1 Descripteur SIFT3D

Dans l'optique d'étendre l'utilisation du descripteur SIFT au domaine spatio-temporel, Scovanner et al. en 2007 [99] puis Klaser et al. en 2008 [61] y ajoutent un modèle d'analyse 3D. La figure 4.5 illustre leur descripteur HOG3D (Klaser et al.), en détaillant les différentes étapes de construction.

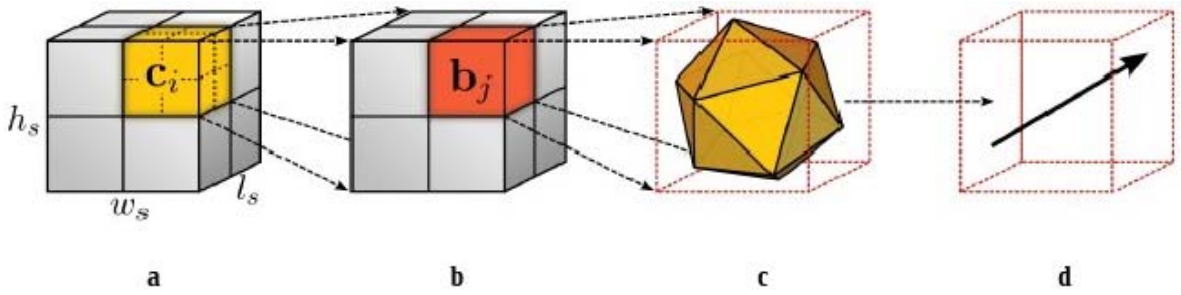


FIG. 4.5 – Représentation des différentes étapes de la construction du descripteur HOG3D : échantillonage du masque d'analyse (a et b), détermination de l'orientation du gradient (d) dans chaque sous-bloc à l'aide d'un icosaèdre (c). (images extraites de [61])

L'approche consiste à déterminer une région d'analyse 3D de taille $h_s \times w_s \times l_s$ centrée sur le point d'intérêt. Le masque d'analyse est alors divisé en M×M×N blocs $\mathbf{c}_i$ (figure 4.5.a) de taille identique qui seront à leur tour divisés en S^3 sous-blocs $\mathbf{b}_j$ (figure 4.5.b). Dans chacun des sous-blocs, l'orientation du gradient (figure 4.5.d) est déterminée à l'aide d'un polyèdre régulier (figure 4.5.c). Dès lors, un histogramme de gradients orientés est construit dans chaque $\mathbf{b}_j$. Les différents paramètres utilisés dans cette approche (h_s, w_s, l_s, M, N et S) sont définis suivant un certain nombre de contraintes. Les dimensions du masque d'analyse sont égales à :

$$w_s = h_s = \sigma_0 \sigma_s \qquad et \qquad l_s = \tau_0 \tau_s, \tag{4.8}$$

où σ_s et τ_s sont les échelles locales extraites par le détecteur de points d'intérêt. Les paramètres σ_0 et τ_0 sont quant à eux les paramètres spatiaux et temporels de la région d'intérêt. Les auteurs préconisent les valeurs suivantes : $\sigma_0 = 8$ et $\tau_0 = 6$. Les paramètres de division de blocs sont déterminés suivant deux critères : le temps de calcul et les performances du descripteur. Le but est de trouver le meilleur compromis. Klaser et al. concluent avec M=4, N=4 et S=3 comme étant les valeurs optimales.

En s'appuyant sur les travaux de Dollar et al. [38], Cheung et Hamarneh proposent en 2009 [23] une amélioration de l'approche précédemment détaillée. L'idée générale est d'utiliser un cube de taille fixe 4×4×4 pour décrire le voisinage du point d'intérêt. La figure 4.6 présente le masque d'analyse du descripteur et détaille la composition d'un des voxels[1].

[1]contraction de "*volumetric pixel*" correspond à un pixel en 3D

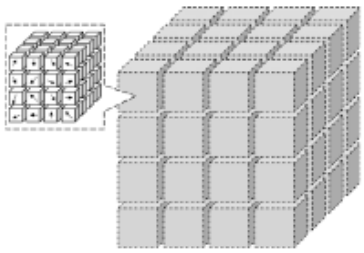


FIG. 4.6 – Masque d'analyse du descripteur n-SIFT et illustration de la composition d'un des voxels. (images extraites de [23])

Chaque voxel est à son tour divisé en soixante-quatre sous-régions (cube 4×4×4) dont chaque case représente l'orientation et la norme du gradient local. Ce descripteur est généralement utilisé dans des analyses tridimensionnelles, néanmoins les expérimentations ont prouvé que son utilisation au domaine spatio-temporel était possible.

4.1.2.2 Généralisation du descripteur SURF

Une extension spatio-temporelle du SURF a été proposée en 2008 par Willems et al. [111]. Le principe est d'étendre les ondelettes de Haar à un cuboïde de taille $s\sigma \times s\sigma \times s\tau$, où σ et τ sont respectivement l'échelle spatiale et l'échelle temporelle et s un facteur défini de façon arbitraire par l'utilisateur. A l'instar des deux descripteurs précédemment cités, la région d'analyse est divisée en M×M×N blocs, possédant chacun l'information $\mathbf{v}$ extraite des ondelettes et définie par :

$$\mathbf{v} = (\sum d_x, \sum d_y, \sum d_t), \tag{4.9}$$

où d_x,d_y et d_t représentent respectivement la réponse des ondelettes de Haar en x, y et t. Nous retrouvons notamment ce descripteur dans des processus de classification ou de synchronisation de séquences vidéo.

4.1.2.3 Descripteur utilisant un couplage de HOG et de HOF

En 2006 [65] puis en 2007 [63] Laptev et al. combinent différents histogrammes afin de définir au mieux l'aspect spatial et l'aspect temporel. L'idée est de construire un histogramme de gradients orientés à l'aide d'une analyse spatiale 'classique' et de le coupler avec un histogramme de flot optique (HOF) afin d'avoir une notion temporelle de la scène étudiée. Le figure 4.7 schématise cette fusion d'histogrammes.

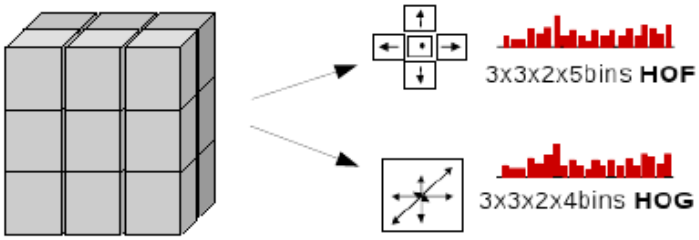


FIG. 4.7 – Illustration du couplage des histogrammes pour la construction du descripteur HOG/HOF (images extraites de [63]).

Tout comme les méthodes précédentes, la taille du masque d'analyse et le nombre de sous-blocs utilisés par cette méthode dépendent des échelles locales du point d'intérêt.

Nous avons présenté différentes approches de détection et de description locales intégrant parfaitement la notion d'analyse temporelle. L'étude de ces méthodes nous permet d'en extraire les principaux avantages (stabilité, performaces et invariances). L'intégration de ces différents outils à notre approche REFA nous permet d'en proposer une généralisation. Par conséquent nous détaillerons dans les paragraphes suivants les modifications engendrées, les nouveaux paramètres utilisés ainsi que les optimisations apportées.

4.2 Méthode REFA3D

La généralisation de notre approche se base sur l'ajout de nouveaux outils présentés précédemment (HOG construit à partir d'un icosaèdre par exemple). Ces derniers doivent être optimisés afin de nous permettre de pallier les problèmes rencontrés au §3.5.7, lors des tests de notre méthode sur des séquences vidéo. Nous allons donc détailler les modifications apportées aux différentes étapes (détection, description, mise en correspondance) de notre approche.

4.2.1 Optimisation de l'extraction des primitives

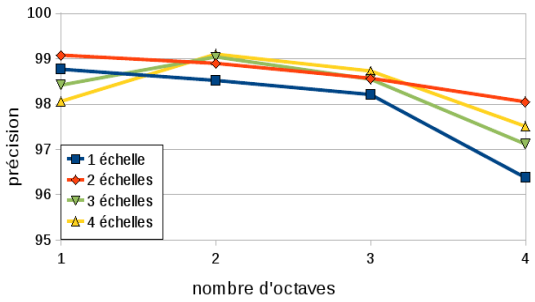
Une analyse du taux de répétabilité des différentes méthodes de détection spatio-temporelle nous permet de déterminer celle présentant le meilleur résultat. Dans la littérature, assez peu d'articles font état d'une telle comparaison. Par conséquent, nous nous appuyons essentiellement sur les résultats présentés dans les articles de Wang et al. [110] et de Willems at al. [111]. Il apparaît qu'une généralisation spatio-temporelle du détecteur fast-hessien, nommée hes-STIP (*hessian spatio-temporal interest point*), obtient la meilleure répétabilité. La construction de ce dernier se base sur l'interprétation de la matrice hessienne (équation 4.4) s'appuyant elle-même sur deux échelles locales σ et τ. La première correspond à l'exploration spatiale définie par le fast-hessien et la seconde permet d'ajouter une analyse temporelle de l'information. Willems et al.

préconisent l'utilisation de quatre octaves pour l'exploration spatiale couplée à quatre valeurs d'échelle d'analyse temporelle. Dans une optique d'optimisation du détecteur hes-STIP nous analysons l'influence de ces deux échelles sur les performances de notre approche. La figure 4.8 présente les résultats obtenus par notre méthode pour une transformation de type changements d'échelle temporelle. La construction des différentes séquences utilisées s'appuie sur la sélection d'une image sur deux, une sur trois, une sur quatre,..., issues de la séquence initiale.

nombre d'échelles temporelles

Nombre d'octaves \	1	2	3	4
1	98,76	99,07	98,41	98,05
2	98,51	98,89	99,03	**99,09**
3	98,2	98,56	98,54	98,72
4	96,37	98,04	97,11	97,5

Précision



nombre d'échelles temporelles

Nombre d'octaves \	1	2	3	4
1	352	534	633	691
2	485	681	723	**766**
3	498	701	759	795
4	507	724	782	823

Nombre d'appariements

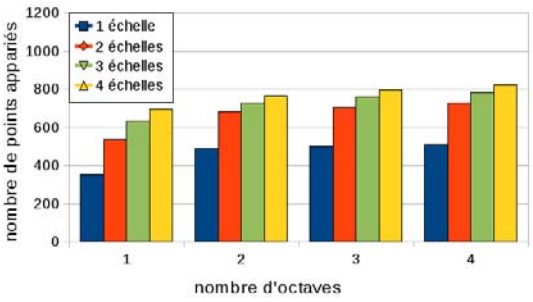


FIG. 4.8 – Analyse de l'influence des espaces d'échelles sur la précision (haut) et le nombre de points appariés (bas), pour une transformation synthétique de type changements d'échelle temporelle.

Les résultats montrent que la précision optimale est obtenue pour une analyse spatiale suivant deux octaves et temporelle suivant quatres échelles. Concernant le nombre de points mis en correspondance, le maximum est atteint pour des valeurs égales à quatre octaves et quatre échelles. Les applications auxquelles notre méthode est destinée s'appuient essentiellement sur la qualité des couples de points, par conséquent la précision est prioritaire sur le nombre de points appariés. Prenons par exemple le cas d'une estimation de la matrice d'homographie, une perte de 7% des points est négligeable par rapport à une augmentation de 1,59% de la précision. En effet, le nombre de points n'est pas (ou peu) influent car une telle estimation peut être réalisée avec un nombre réduit de couples. En contrepartie, la justesse des appariements accroît fortement la qualité et la précision de l'estimation, les faux appariements (*outliers*) étant moins nombreux. En définitive, nous optons pour une analyse spatiale limitée à deux octaves et l'utilisation de quatre échelles temporelles.

4.2.2 Modification de la description locale

Nous avons détaillé au §3.2 le processus de description locale de notre approche. Nous souhaitons généraliser ce procédé afin d'y intégrer des données temporelles. Néanmoins, cet ajout entraîne des modifications : le masque d'analyse, l'angle de recalage ou la construction des histogrammes. Par conséquent, nous proposons d'en détailler les différents paramètres nous permettant ainsi de proposer une description spatio-temporelle du point d'intérêt.

4.2.2.1 Masque d'analyse

La description locale de la méthode REFA s'appuie sur l'utilisation d'histogrammes de gradients orientés suivant un masque elliptique (figure 3.7). L'ajout de données temporelles nous oblige à modifier notre masque, transformant ainsi les ellipses en ellipsoïdes. Afin d'analyser l'intégralité de l'information spatio-temporelle, nous proposons donc le masque présenté en figure 4.9, s'appuyant sur un échantillonage ellipsoïdique du voisinage du point d'intérêt. Ce dernier est déterminé suivant cinq niveaux de description (niveau -2 au niveau 2) cumulant 37 ellipsoïdes. Pour une meilleure visibilité de l'aspect spatio-temporel de notre descripteur, nous n'affichons que les centres des ellipsoïdes sur le schéma.

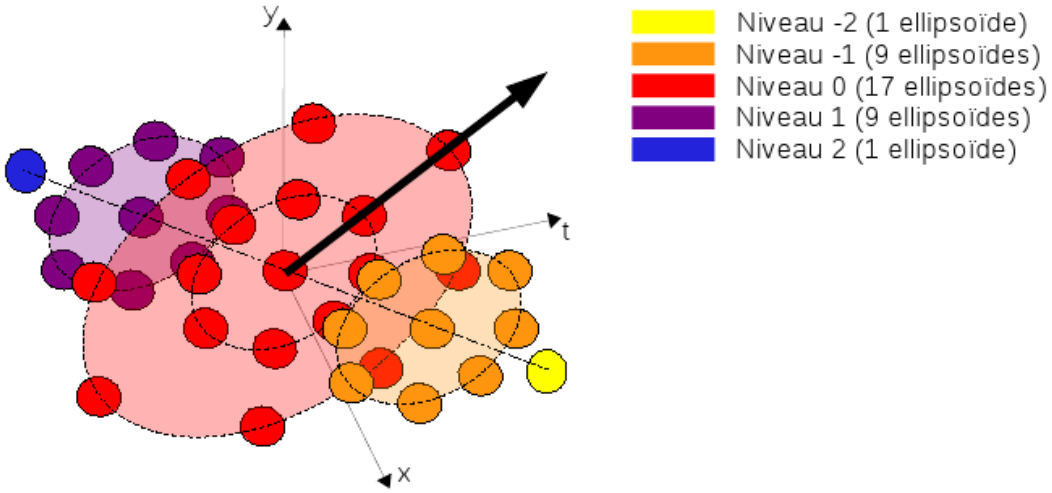


FIG. 4.9 – Représentation de notre masque d'analyse ellipsoïdique, suivant cinq niveaux de description.

La figure 4.10 présente quant à elle une vue détaillée du niveau central de description (niveau 0) et donnant un aperçu des paramètres nécessaires à la construction de chaque ellipsoïde (σ_1, σ_2 et τ). Les paramètres des ellipsoïdes se basent sur les échelles locales des points d'intérêt. La détermination du rapport entre σ_1 et σ_2 est conservée. La valeur du paramètre τ est égale à l'échelle temporelle retournée par le détecteur. Cette dernière étant optimisée, aucune étape supplémentaire (ajustement, seuillage) n'est donc nécessaire.

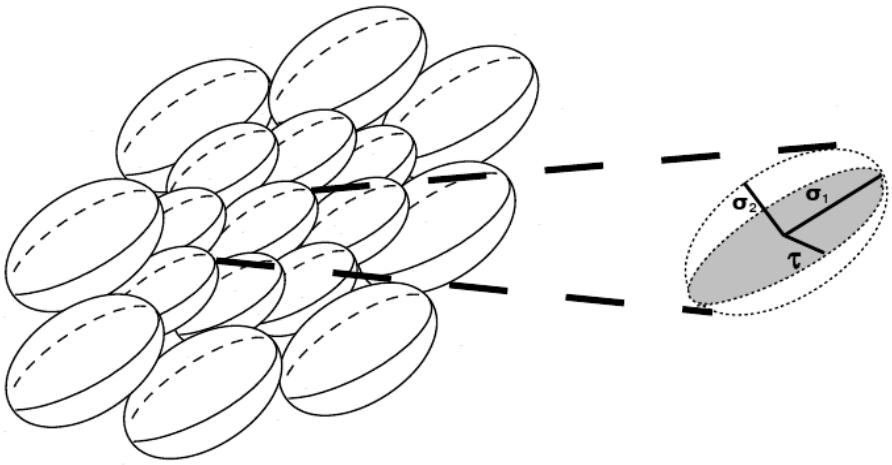


FIG. 4.10 – Représentation du niveau central de description. Chaque ellipsoïde est construite à l'aide des paramètres σ_1, σ_2 et τ issus du détecteur.

4.2.2.2 Détermination de l'angle de recalage

Nous avons également proposé au §3.2.2 une étape de traitement préliminaire permettant d'accroître l'invariance à la rotation en recalant le masque d'analyse. Ce recalage s'appuie sur l'utilisation d'un angle θ extrait de l'interprétation de la matrice de Harris (équation 2.17). Son utilisation est étendue afin d'en extraire les composantes d'un recalage spatio-temporel. L'analyse de la matrice de Harris3D (équation 4.1) introduite par Laptev et Lindeberg permet de récupérer deux angles θ et φ, présentés en figure 4.11. Cette dernière schématise, d'un coté le recalage spatial, et de l'autre le recalage temporel.

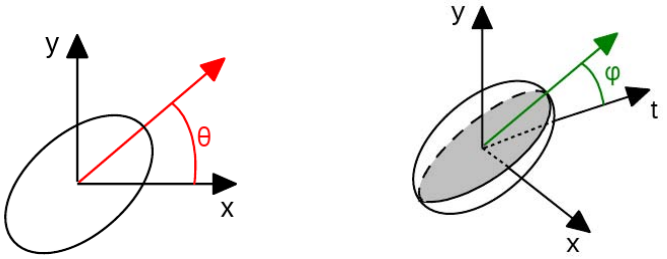


FIG. 4.11 – Illustration du recalage spatial (à gauche) et temporel (à droite) d'une ellipsoïde constituant notre masque d'analyse.

4.2.2.3 Optimisation de la construction du descripteur

La construction de notre descripteur repose essentiellement sur l'utilisation d'histogrammes de gradients orientés constitués de huit classes. L'ajout de données temporelles

nous oblige donc à modifier ces histogrammes. Nous appuyant sur les travaux de Klaser et al. [61], présentant une généralisation des HOG au domaine spatio-temporel, nous les construisons suivant vingt classes. Pour ce faire, notre descripteur se base sur un icosaèdre (polyèdre régulier préconisé par Klaser et al.) permettant notamment de simplifier la répartition des données. Le choix de la classe de l'histogramme repose sur la détermination de l'intersection du vecteur gradient avec une des vingt faces de l'icosaèdre. Ceci permet de récupérer l'indice de la face du polyèdre et d'en déduire la classe d'appartenance du point analysé. La figure 4.12 schématise l'icosaèdre utilisé lors de la construction de nos HOG spatio-temporels.

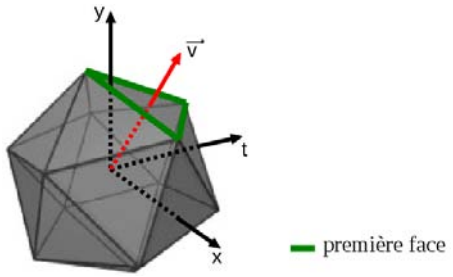


FIG. 4.12 – Représentation de l'icosaèdre utilisé pour la construction de nos HOG3D.

Afin d'ordonner notre descripteur de façon optimale, la face représentant la première classe de nos histogrammes est recalée suivant le vecteur **v**. Ce dernier correspond à la combinaison des recalages représentés en figure 4.11 (spatial suivant θ et temporel suivant φ). Ce procédé classe de façon identique les données provenant de deux points devant s'apparier.

De plus une étape de saturation, détaillée au §3.2.3, accroît la robustesse aux changements d'illumination. Pour ce faire, nous nous appuyons sur l'analyse présentée par la figure 4.13 regroupant les performances de notre approche REFA3D en fonction de la valeur du seuil.

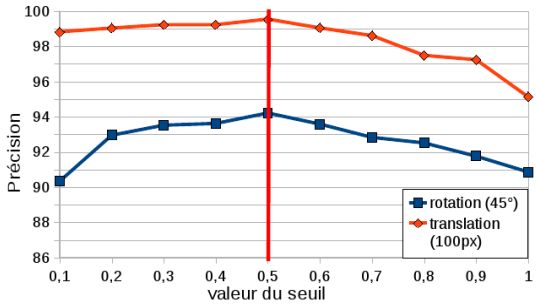


FIG. 4.13 – Détermination du seuil de saturation des HOG spatio-temporels pour deux types de transformations : une rotation de 45°et une translation de 100 pixels.

Nous observons qu'une valeur de seuil de saturation égale à 0,5 permet d'optimiser la précision de notre méthode. Ce seuil permet entre autre de limiter l'influence de données aberrantes caractérisées par de fortes valeurs de gradient.

4.2.3 Mise en correspondance

Nous souhaitons conserver la méthode de mise en correspondance détaillée au §3.3. Cette dernière se base, d'une part sur une maximisation de la corrélation entre les descripteurs locaux, et d'autre part sur des optimisations (seuil de sélection, suppression des doublons) qui accroient les performances. Nous généralisons donc cette approche en intégrant la prise en compte des données temporelles lors de la construction de l'arbre de décision (détaillé au §3.3.1). Ce dernier dépendant de la taille des données fournies, sa dimension est donc de $\mathbb{R}^{740}$ (trente-sept histogrammes de vingt classes chacun). Concernant le seuil de sélection et la méthode de suppression des doublons, leurs procédés et paramètres restent inchangés. La figure 4.14 illustre l'influence du seuil de sélection sur les performances de notre approche.

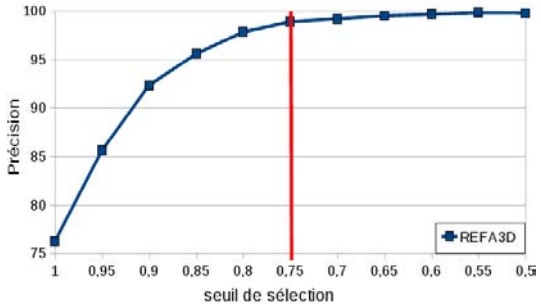


FIG. 4.14 – Détermination du seuil de sélection pour un changement d'échelle spatiale (Pavin1, voir annexe B).

La valeur $0,75$ représente pour le coefficient α le seuil de stabilité pour lequel les performances cessent de croître de façon significative. Nous optons pour cette valeur, fournissant le meilleur compromis entre la performance et le nombre d'appariements.

4.3 Validation de notre approche

Afin de valider notre généralisation spatio-temporelle, nous proposons deux études basées, d'une part sur des transformations synthétiques, et d'autre part sur des transformations réelles. Pour ce faire, nous utilisons la plate-forme PAVIN présentée en annexe B afin d'en extraire des séquences présentant différentes approches. Nous testons notamment les translations, les rotations et nous complétons cette validation par une simulation de l'accélération du déplacement de la caméra. Pour obtenir ce type de transformation nous récupérons une image sur deux, une sur trois, une sur quatre, ... de la séquence initiale. Notre étude se porte principalement sur la précision et le taux d'appariements obtenus par notre approche. Afin de valider au mieux notre approche, nous

la comparons aux méthodes HOG/HOF (décrite au §4.1.2.3) et SIFT3D (décrite au §4.1.2.1). Notre principal objectif est de proposer de meilleurs résultats en terme de précision (ayant une répercution directe sur des applications utilisant des couples de points). D'autre part nous avons vu au chapitre précédent que le nombre de points mis en correspondance dépendait fortement du nombre de points détectés, nous préconisons donc d'étudier sa stabilité (basée sur la robustesse de la description) ainsi que son taux d'appariement.

4.3.1 Transformations synthétiques (la vérité terrain)

Les transformations synthétiques étudiées se divisent en différentes parties : la translation permettant de tester la description locale, la rotation mettant à défaut les différents paramètres de recalage, les changements d'échelles temporelles et spatiales étudiant les descriptions de manière plus précise et validant les échelles locales extraites. Ce type de transformations sera appliqué sur les séquences 1 et 2 issues de la plate-forme PAVIN (détaillées en annexe B). La figure 4.15 donne un aperçu des transformations étudiées.

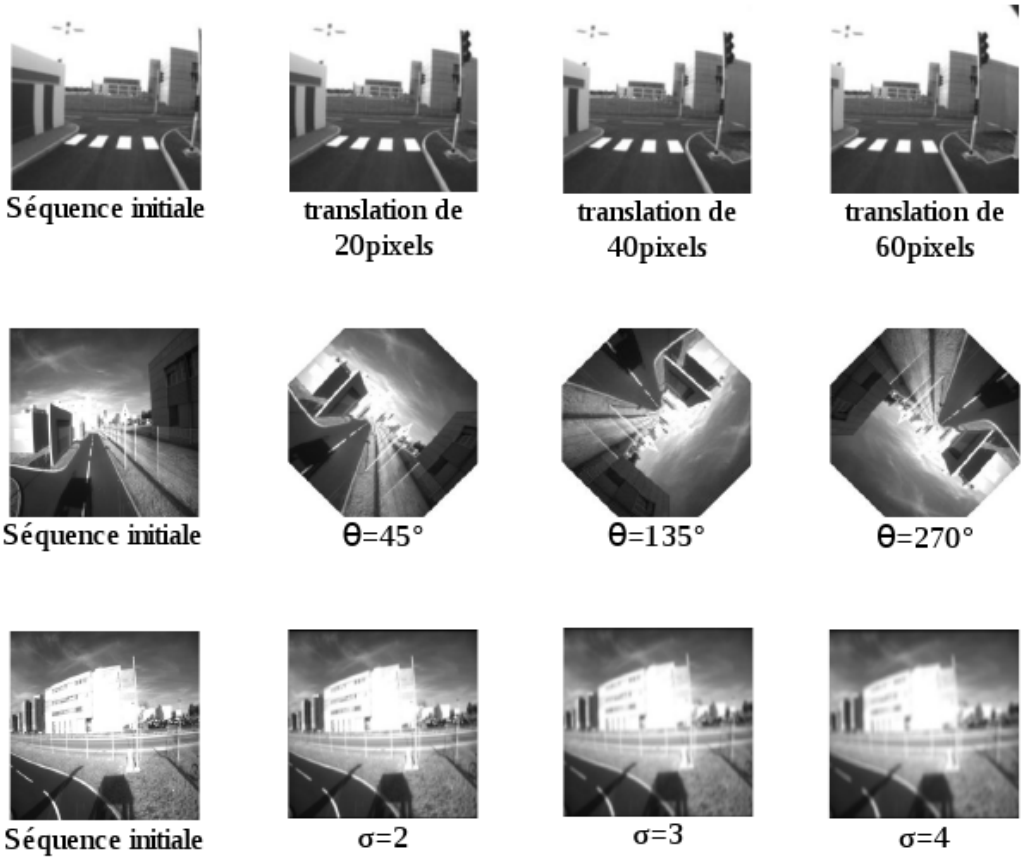


Fig. 4.15 – Exemples de transformations (translations, rotations et changements d'échelle spatiale) basées sur les images PAVIN.

– Nous souhaitons d'analyser des transformations de type **translations**. Les figures F.19 et F.20 illustrent la précision obtenue et le nombre de points mis en correspondance pour les séquences 1 et 2.

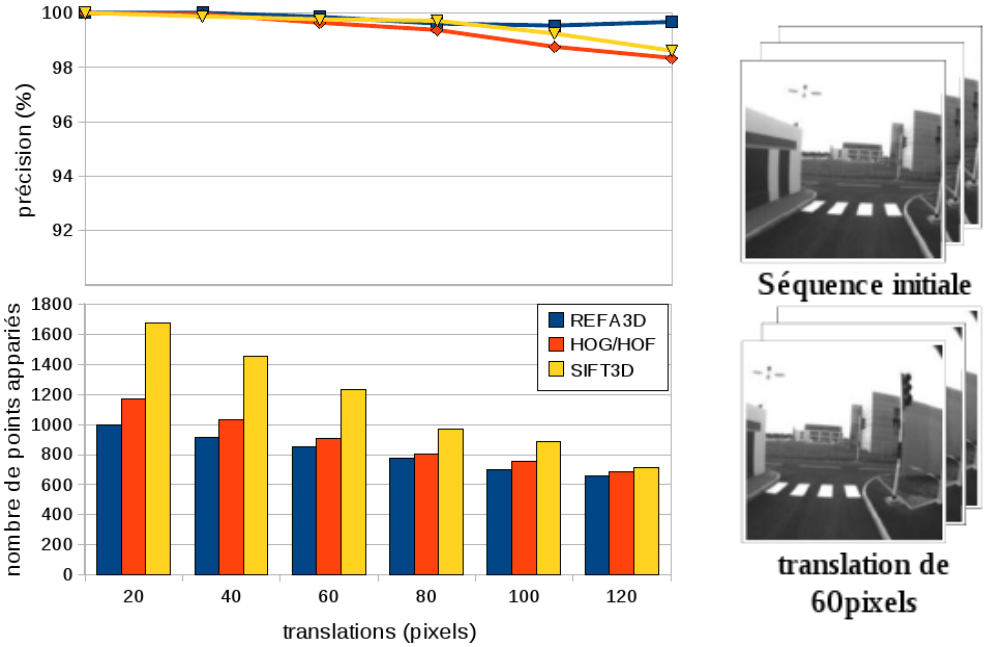


FIG. 4.16 – Précision et nombre de points appariés pour des translations horizontales (séquence 1).

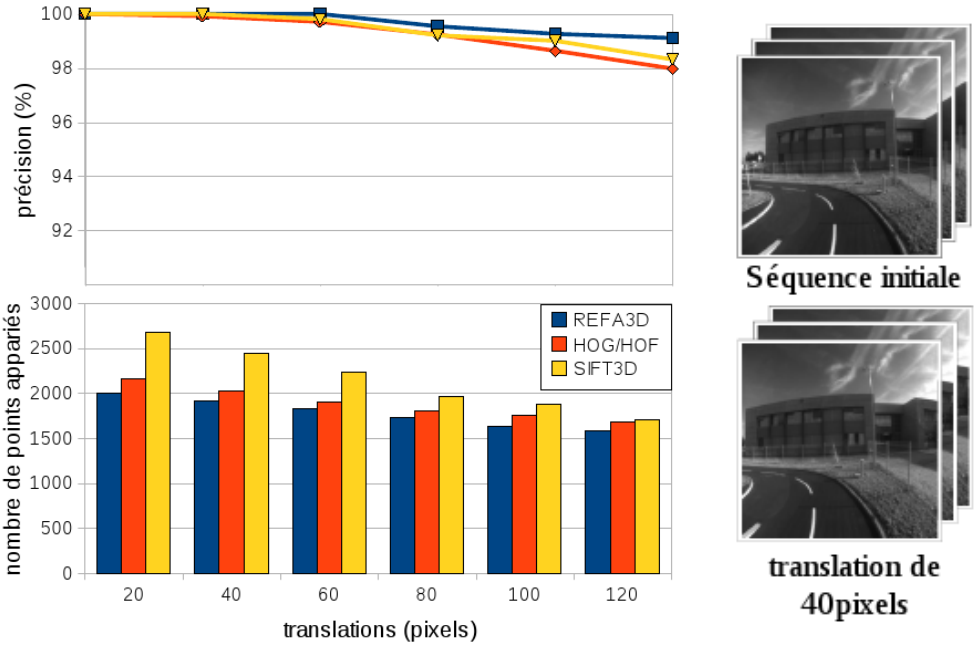


FIG. 4.17 – Précision et nombre de points appariés pour des translations horizontales (séquence 2).

Pour les translations notre méthode présente la meilleure précision. Nous obtenons également une meilleure stabilité en terme de nombre de points appariés du fait de la décroissance moins rapide de ce dernier.

– Nous proposons également d'étudier l'influence de modifications de type **rotations** sur les performances de notre méthode. La figure F.21 regroupe les résultats obtenus, précision et nombre de points appariés, pour la séquence 2.

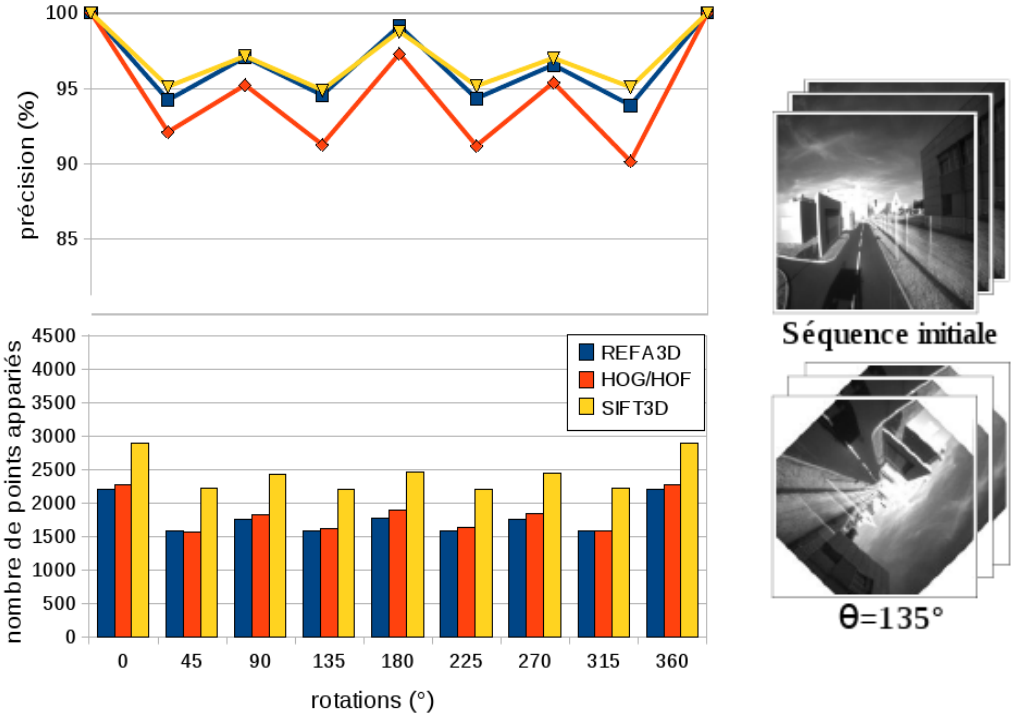


FIG. 4.18 – Précision et nombre de points appariés pour des rotations (séquence 2).

Concernant les rotations, nous avons détaillé au §4.1.2.1 que le SIFT3D s'appuie sur le SIFT [73][74] lui procurant ainsi une parfaite gestion de ce type de transformations. En utilisant un recalage spatial sur notre masque de description locale, nous obtenons des résultats similaires au SIFT3D pour les rotations. D'autre part, le nombre de points mis en correspondance est plus faible car il y a moins de points initialement détectés. Cependant sa stabilité reste similaire à celle des deux autres méthodes.

– L'étude de **changements d'échelle spatiale** permet de mettre en avant l'utilité de l'adaptation des ellipsoïdes et de la qualité du paramètre σ extrait (figure 4.10). Les figures F.22 et F.23 regroupent les résultats obtenus pour ce type de transformations.

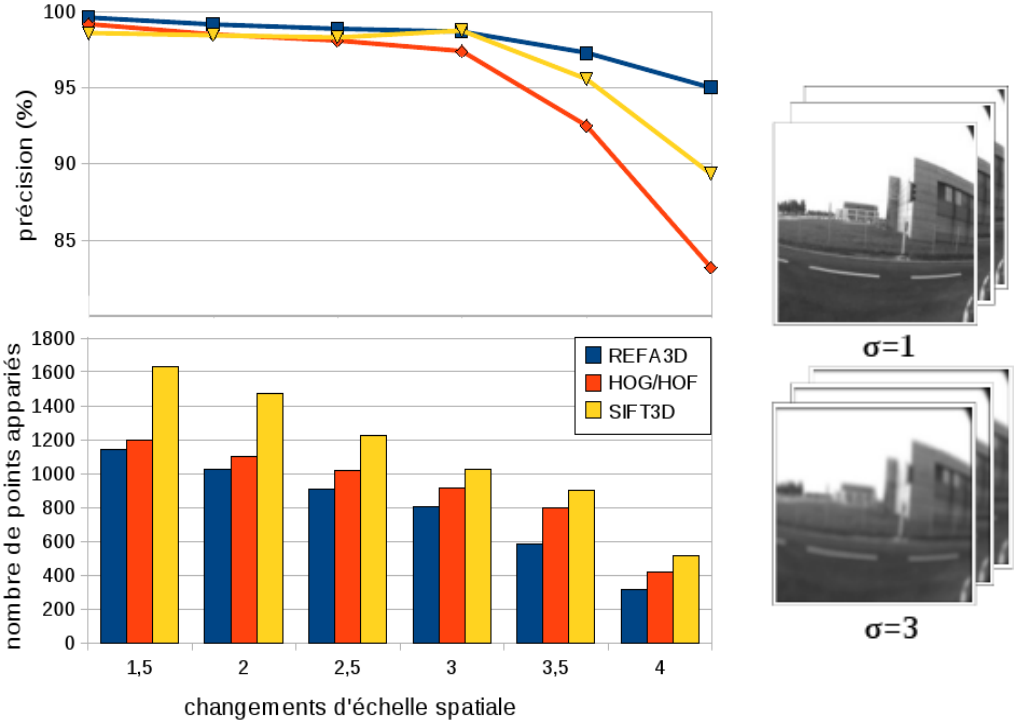


FIG. 4.19 – Résultats pour des changements d'échelle spatiale (synthétique ; séquence 1).

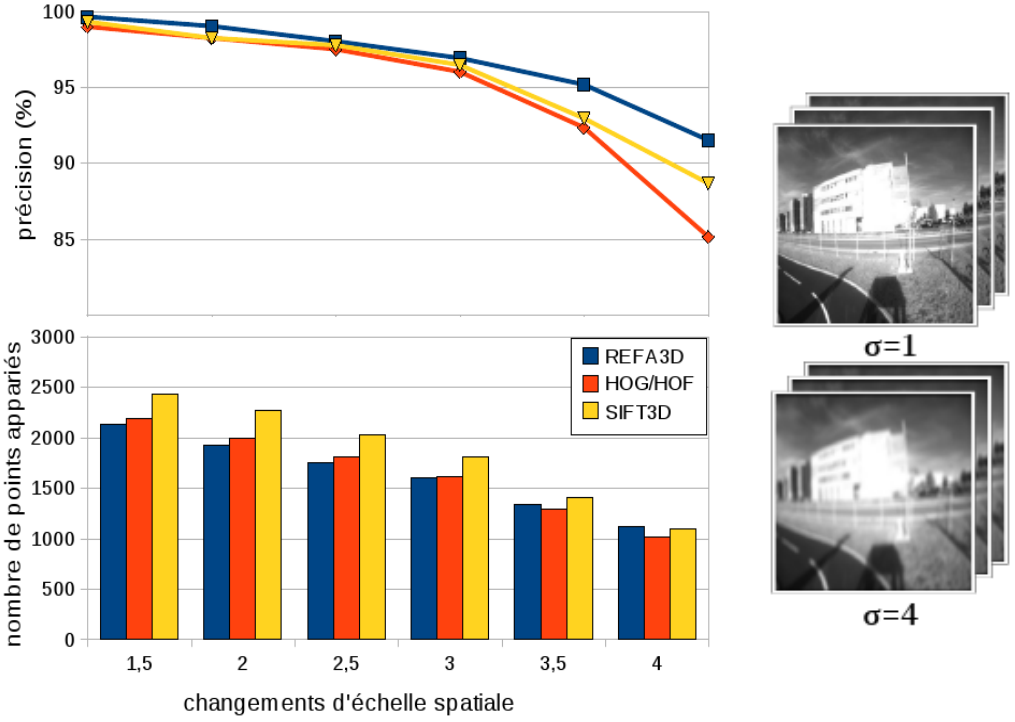


FIG. 4.20 – Résultats pour des changements d'échelle spatiale (synthétique ; séquence 2).

Notre approche REFA3D présente une précision plus élevée notamment pour de fortes modifications d'échelle spatiale ainsi qu'une stabilité plus prononcée. En effet, d'une part sa précision diminue moins rapidement que celles du HOF/HOF et du SIFT3D, et d'autre part le nombre de points que nous apparions décroît plus lentement. La mise à l'échelle de notre masque ellipsoïdique permet donc d'obtenir une meilleure description locale du voisinage du point pour ce type de transformations. Par conséquent les paramètres σ_1 et σ_2 (définis en figure 4.10) jouent un rôle important dans notre approche et leur détermination influence nos performances.

– Nous proposons une dernière étude simulant une accélération du déplacement de la caméra. Les **changements d'échelles temporelles** nous permettent de mettre en avant, d'une part l'importance des données temporelles utilisées lors de la description locale, et d'autre part la qualité du paramètre τ extrait (figure 4.10). Les figures F.24 et F.25 illustrent les résultats pour ce type de transformations. Les absisces correspondent aux intervalles de sélection des images prélevées dans la séquences initiales (2 : 1 image sur 2, 3 : 1 image sur 3, ...).

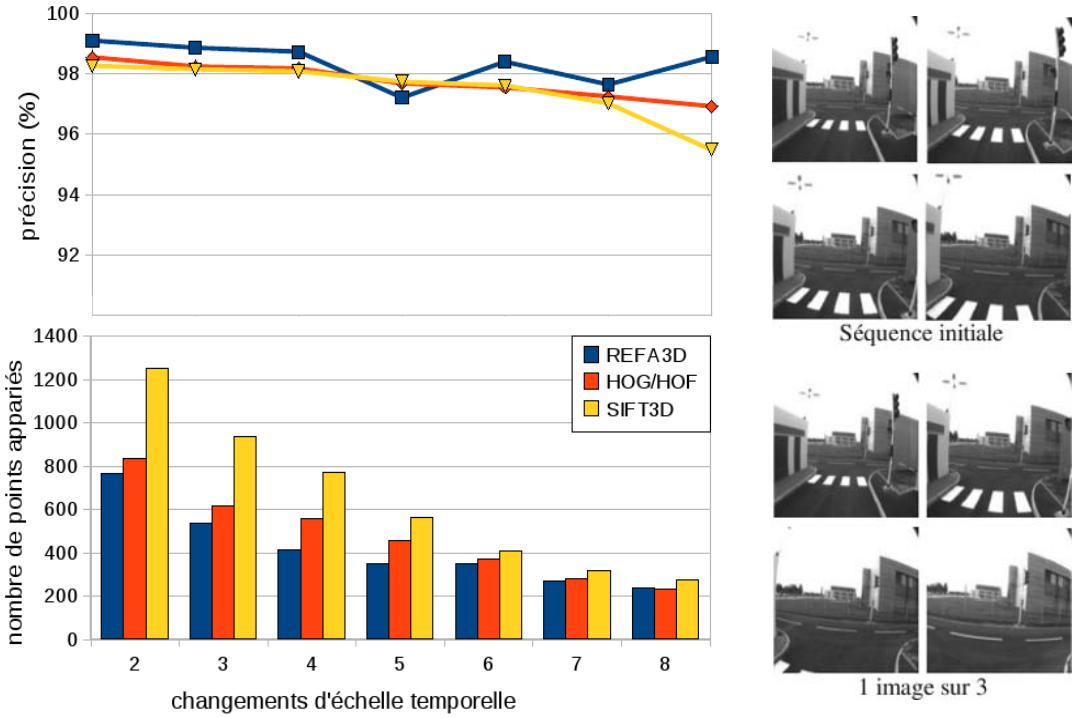


FIG. 4.21 – Résultats pour des changements d'échelle temporelle (synthétique ; séquence 1).

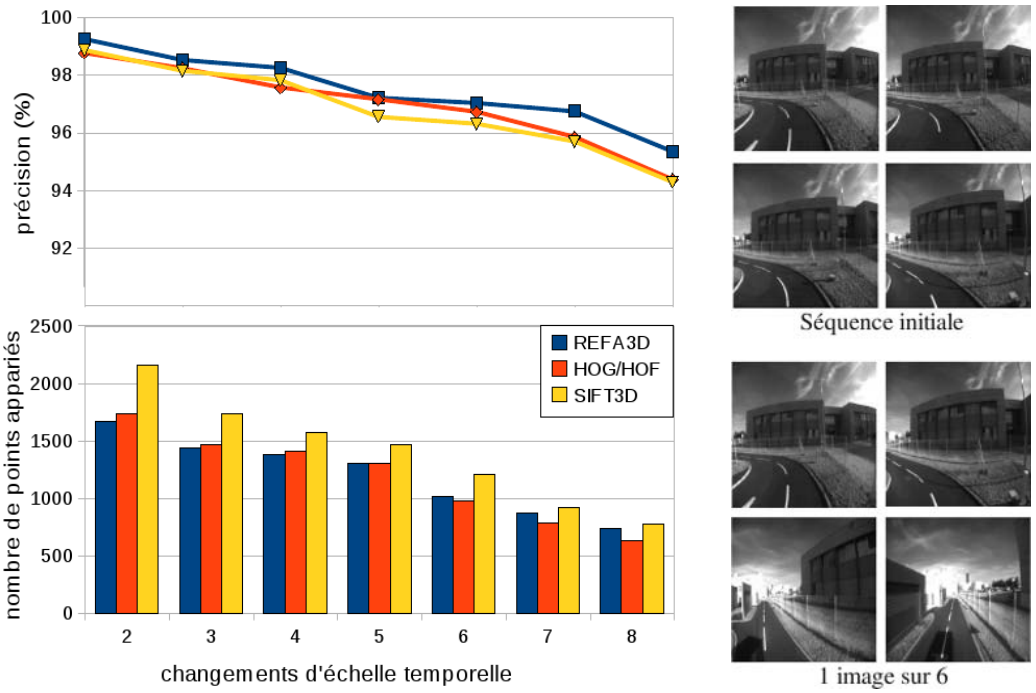


FIG. 4.22 – Résultats pour des changements d'échelle temporelle (synthétique ; séquence 2).

Pour des transformations de type changements d'échelle temporelle, notre méthode présente la meilleure précision et est également plus stable. L'ajout de données temporelles et le recalage qui lui est associé permet donc d'accroître la robustesse à d'éventuelles modifications de vitesse de déplacement de la caméra.

En définitive, concernant les transformations synthétiques étudiées, notre approche présente une précision plus élevée que celles des autres méthodes ou similaire au SIFT3D dans le cas de rotations. La stabilité nous permet également de conclure à une meilleure robustesse de notre approche. Néanmoins, ces différents résultats reposent sur divers outils (optimisations, seuils) impliquant une diminution du nombre de points appariés. Cette différence, importante vis à vis du SIFT3D, est principalement due aux nombres de points extraits par le détecteur. En effet, une méthode telle que SIFT3D fournit un nombre de points initial plus élevé. Afin d'analyser de façon plus judicieuse ces différentes valeurs, nous utiliserons au §4.3.4 le taux d'appariements défini précédemment.

4.3.2 Transformations réelles

Les premières transformations réelles proposées reposent sur les séquences extraites du simulateur ASROCAM présentées en annexe D. Nous comparons notre approche aux SIFT3D et HOG/HOF suivant deux jeux de séquences. Le premier correspond à un quart de tour du rond-point de la plateforme PAVIN et le second s'appuie quant à lui sur le tour complet. Chaque jeu se compose de trois trajectoires (intérieure, centrée et extérieure) étudiées les unes par rapport aux autres et mises en correspondance. Ces

différents tests permettent de mettre en avant la robustesse des méthodes étudiées vis à vis d'un ensemble de transformations pouvant s'apparenter à une dérive du véhicule. Les figures F.26 et F.27 illustrent les résultats obtenus.

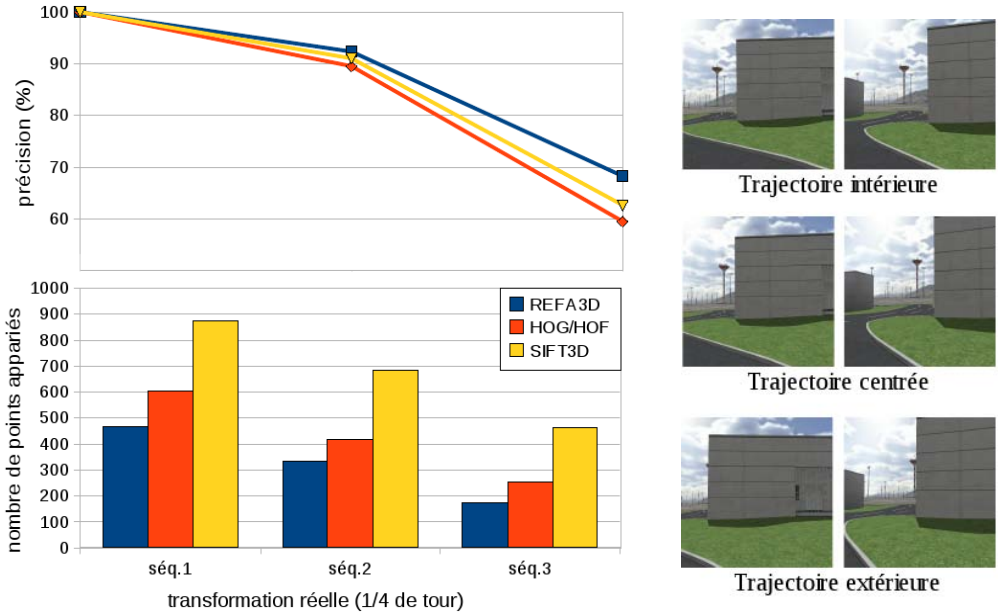


FIG. 4.23 – Résultats des mises en correspondance des 3 trajectoires issues d'un quart de tour du rond-point (ASROCAM).

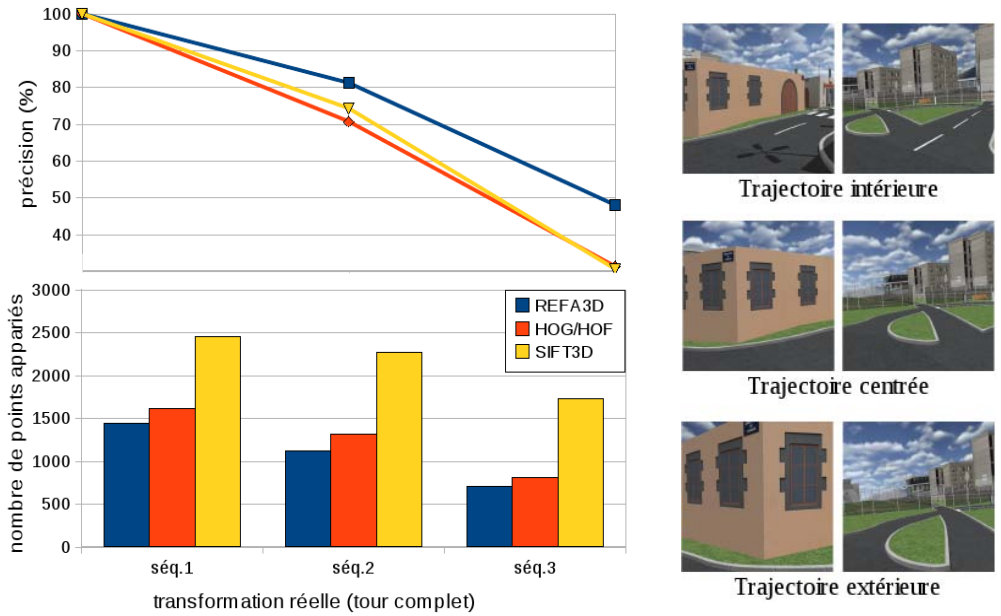


FIG. 4.24 – Résultats des mises en correspondance des 3 trajectoires issues du tour complet du rond-point (ASROCAM).

Les performances obtenues lors de la mise en correspondance de la séquence 1 avec les autres séquences montrent que notre approche est plus robuste que les méthodes SIFT3D et HOG/HOF. En détaillant ces courbes, nous observons des performances similaires entre les trois méthodes pour la mise correspondance de la trajectoire intérieure avec la trajectoire centrale lors du quart de tour. Concernant l'appariement intérieure/extérieure, le véhicule s'écarte sensiblement du centre du rond-point, entraînant une diminution de la précision et du nombre de points mis en correspondance. Notre méthode REFA3D décroit plus lentement, caractérisant ainsi une meilleure stabilité. Les tests effectués sur le tour complet du rond-point mettent en avant l'accumulation des erreurs d'appariements, représentée par une forte diminution de la précision (plus de 20% pour l'appariement trajectoire intérieure/trajectoire extérieure). En conclusion, notre méthode présente donc des performances décroissantes,tout comme SIFT3D et le couplage HOG/HOF. Elle conserve néanmoins une précision plus élevée et une meilleure stabilité.

Les secondes transformations réelles que nous étudions sont issues du simulateur présenté en annexe C. Ce dernier permet d'observer la scène suivant deux caméras (une à l'avant du véhicule et une autre à l'arrière), créant ainsi deux séquences d'images pour une seule trajectoire. Pour nos tests, nous proposons d'effectuer une boucle (figure C.1) afin de passer deux fois au même endroit. Cela nous permet d'extraire deux séquences composées d'images issues de la caméra avant pour le premier passage, puis de la caméra arrière pour le second. Le principal problème réside dans la mise en correspondance 'croisée' des deux séquences (schématisée en figure 4.25). En effet pour ce type de test, les séquences sont temporellement inversées : la première image issue de la caméra avant correspond à la dernière image issue de la caméra arrière.

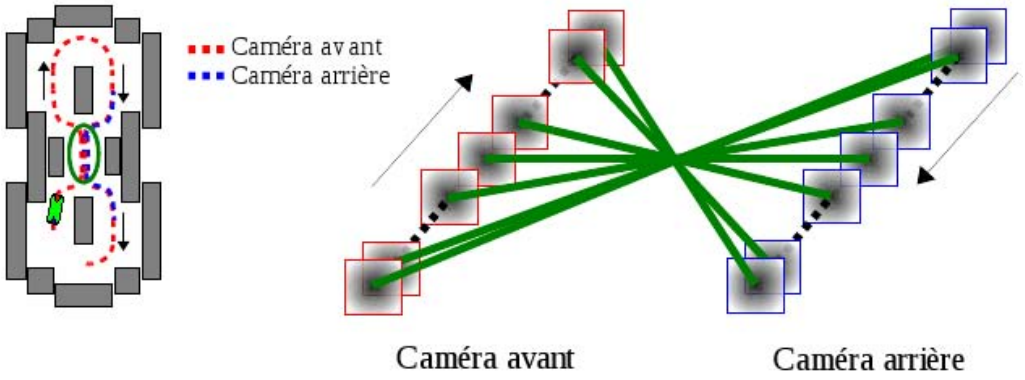


FIG. 4.25 – Représentation d'un appariement "croisé" entre deux séquences issues des caméras avant et arrière du véhicule lors d'une boucle réalisée par simulation (annexe C).

Pour ce type de test, notre approche présente une précision plus élevé (58,7%) que celles du SIFT3D (51,6%) et du couplage HOG/HOF (52,3%). Néanmoins le nombre de points appariés est plus faible (REFA3D : 232 ; SIFT3D : 319 ; HOG/HOF : 253).

Nous expliquons cela par le fait que cette simulation introduit un certain nombre de contraintes : appariement “croisé“, faible quantité de détails dans les images, motifs se répétant.

4.3.3 Influence de la détérioration des données

Afin de valider les différentes aspects de notre méthode, nous étudions sa stabilité vis à vis de la détérioration des données. En effet, dans l'optique d'utiliser notre approche dans des applications de reconstructions 3D, de suivis ou de localisations, cette dernière se doit d'être la plus robuste possible aux perturbations des données. Pour ce faire nous utilisons le procédé décrit au §3.5.4, qui consiste à modifier le seuil de validation α dans le but d'accroître le nombre d'appariements possibles. Les différents tests permettent d'établir les courbes 4.26, 4.27 et 4.28 mettant en avant le *recall* en fonction de 1-précision pour trois séquences issues de la plateforme PAVIN (annexe B).

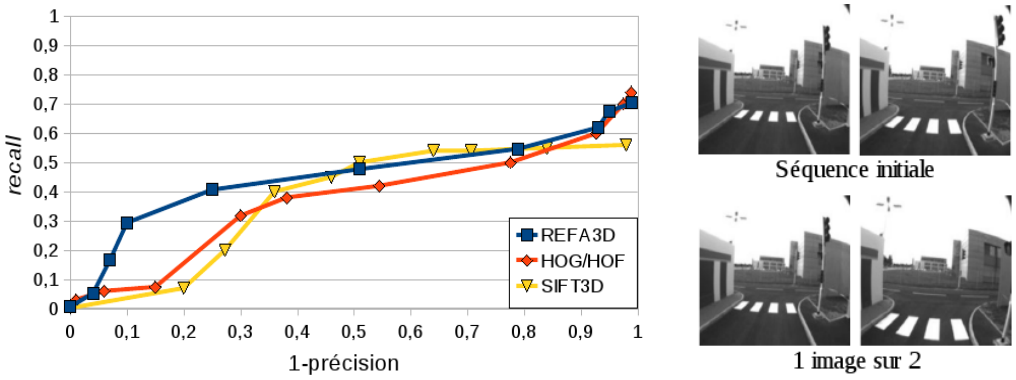


FIG. 4.26 – Etude de l'influence de la dégradation des données pour un changement d'échelle temporelle (séquence 1).

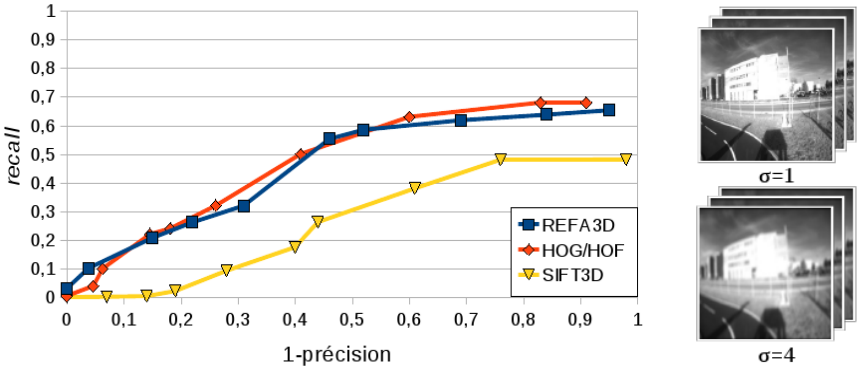


FIG. 4.27 – Etude de l'influence de la dégradation des données pour un changement d'échelle spatiale (séquence 2).

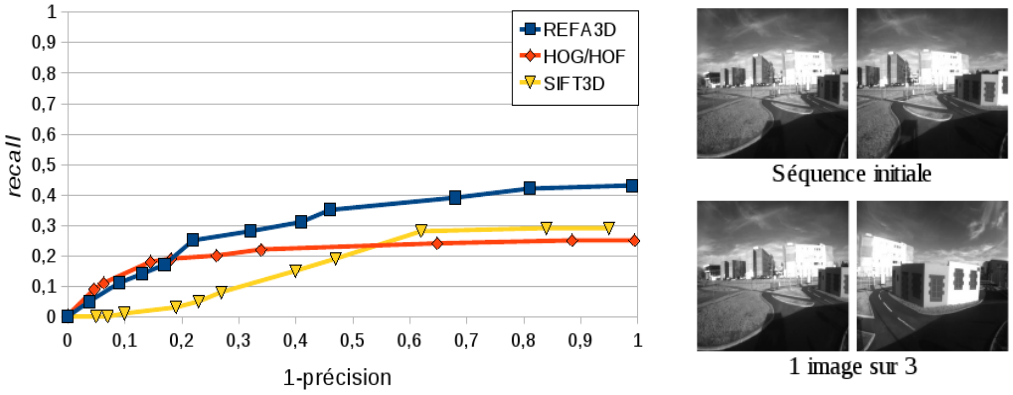
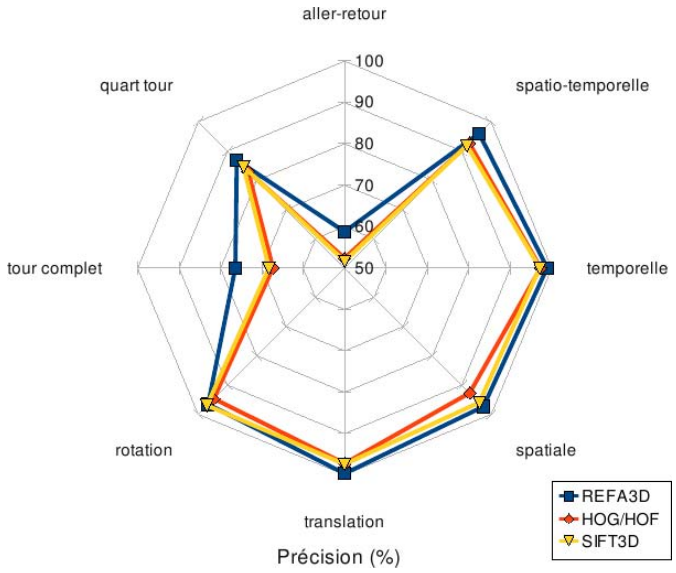


FIG. 4.28 – Etude de l'influence de la dégradation des données pour un changement d'échelle temporelle (séquence 3).

Nous observons d'un point de vue général des performances identiques voir meilleures pour notre approche. Néanmoins les résultats proposés par la couplage HOG/HOF lors de la séquence 2 sont légèrement supérieurs (3% de différence en terme de taux d'appariements corrects). Ceci est la conséquence d'une faible vitesse de déplacement du véhicule et donc d'un grand nombre de points possédant une analyse temporelle très proche. Nous pouvons également voir que les performances proposées par les trois méthodes sont plus faibles pour la séquence 3. En effet, la présence de nombreuses ombres résultant d'un ensoleillement très prononcé peut être assimilée à des changements locaux de luminosité, pénalisant ainsi la description. Couplées à l'ajout croissant de 'faux candidats', les performances s'en retrouvent par conséquent diminuées.

4.3.4 Synthèse des résultats obtenus

Afin de visualiser l'ensemble des résultats présentés dans ce chapitre, nous proposons d'en donner une synthèse. Cette dernière se base sur des regroupements en fonction du type des transformations étudiées. Divers résultats non énoncés dans ce manuscrit, mais nous permettant d'étoffer les performances de notre approche sont ajoutés. Les critères d'observation détaillés au chapitre précédent sont conservés, définissant ainsi la précision et le taux d'appariements. Le nom de chaque regroupement est abrégé de la façon suivante : 'BSS' pour la Base de Séquences ayant subie des transformations Synthétiques ('r' : rotations, 't' : translation, 'es' : changements d'échelle spatiale, 'et' : changements d'échelle temporelle) et 'BSR' pour la Base de Séquences possédant des transformations Réelles ('s' : simulateur présenté en annexe C, 'aq' : simulateur ASROCAM 1/4 de tour, 'at' : simulateur ASROCAM tour complet). La figure 4.29 présente ces deux critères. Elle permet d'avoir une vue d'ensemble des comparaisons effectuées entre notre approche REFA3D, SIFT3D et le couplage HOG/HOF.



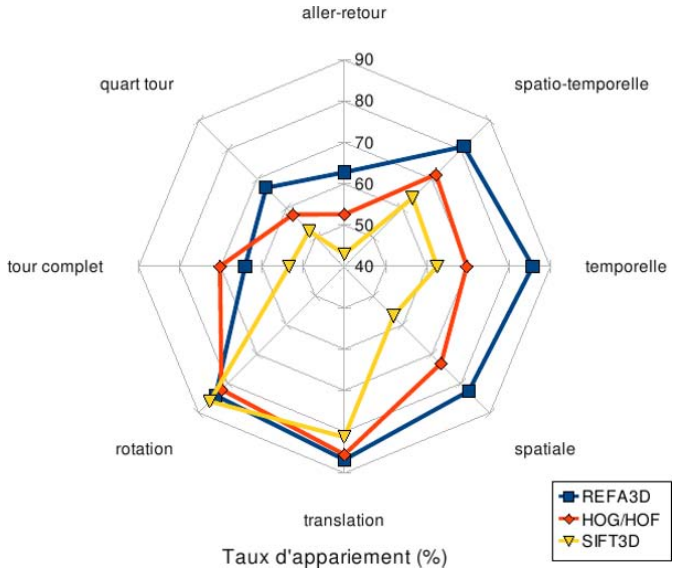


FIG. 4.29 – Synthèse des résultats obtenus dans le domaine spatio-temporel : précision (haut) et taux d'appariement (bas) définis respectivement en 3.6 et 3.7.

Au vu de cette synthèse, il apparaît que notre approche présente les meilleurs résultats dans la majorité des cas. Sa précision décroît lors de transformations réelles, mais reste supérieure à celle du HOG/HOF et celle du SIFT3D. Notre approche présente également un taux d'appariement globalement meilleur, caractérisant une description plus pertinente du voisinage. En définitive notre méthode REFA3D est plus robuste et plus stable vis à vis des différentes transformations étudiées.

4.3.5 Exemple d'application : recalage de sous-séquences

Nous appliquons notre méthode REFA3D sur différentes sous-séquences afin d'en extraire leur localisation dans la séquence d'origine. Le principe consiste à décrire dans un premier temps la totalité de la séquence afin d'avoir une base de comparaison. Dans un second temps, la sous-séquence que nous souhaitons recaler est analysée puis comparée avec la base initialement créée. Au final nous retenons, pour chaque image de la sous-séquence, la valeur d'appariement la plus élevée. Le schéma de la figure 4.30 illustre le principe de recalage d'une sous-séquence.

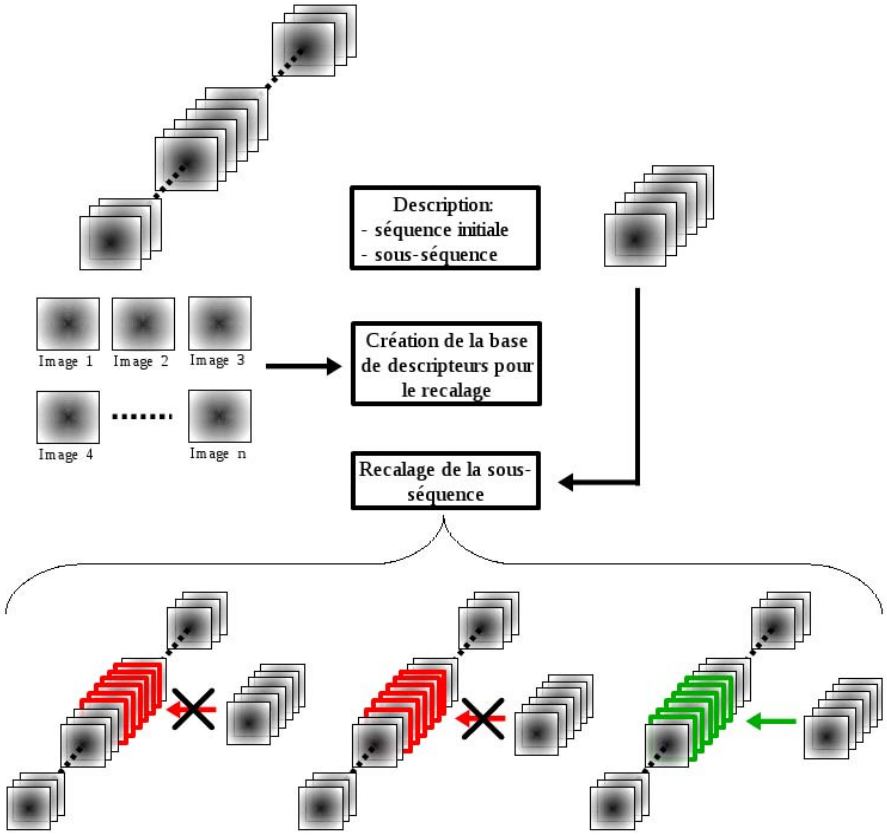


FIG. 4.30 – Schéma illustrant le recalage d'une sous-séquence dans une séquence d'origine. La localisation se base sur une comparaison itérative entre la sous-séquence et la base de descripteurs.

Un exemple de recalage d'une sous-séquence, extraite de la plateforme PAVIN (séquence 1) est présenté en figure 4.31.

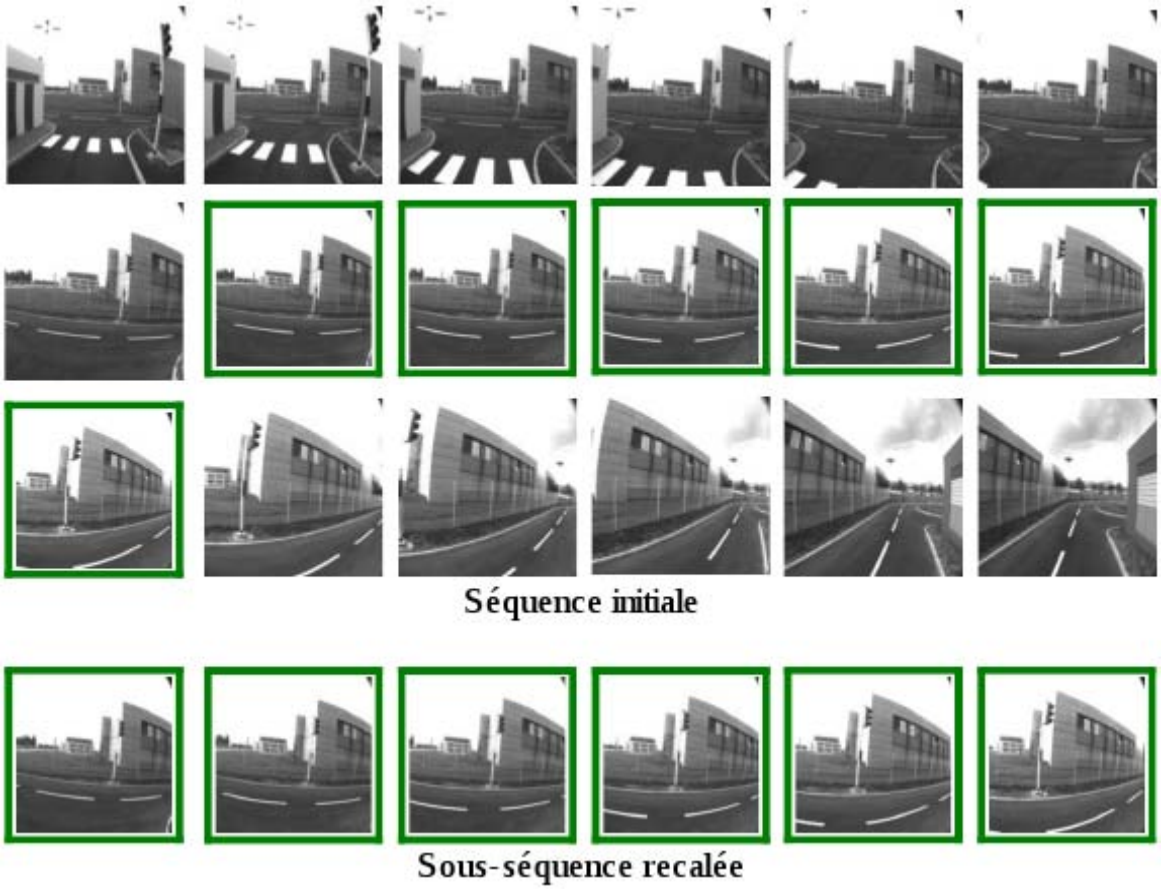


FIG. 4.31 – Exemple de recalage d'une sous-séquence, extraite de la plateforme PAVIN, dans la séquence d'origine.

Afin d'observer l'amélioration des résultats entre une analyse spatiale (résultats présentés en fin de chapitre 3) et une analyse spatio-temporelle, nous proposons d'effectuer un recalage d'une sous-séquence de 10 images dans la séquence 1 de la plateforme PAVIN. Les paramètres de test et les critères d'observation restent inchangés. La figure 4.32 présente la précision d'appariements de notre méthode pour ce type de procédé.

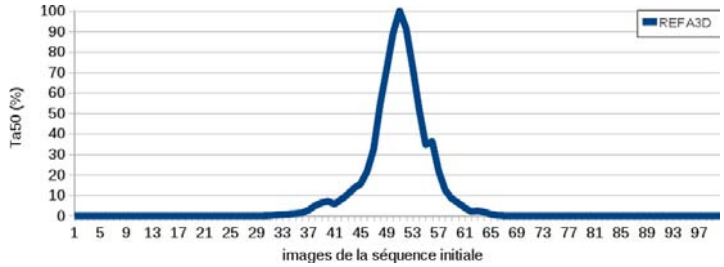


FIG. 4.32 – Précision d'appariements pour le recalage d'une sous-séquence : avantage d'une analyse spatio-temporelle.

Il en résulte que l'ajout de données temporelles entraîne une description plus pertinente, et donc un recalage optimal et précis de la sous-séquence.

Nous testons également différentes sous-séquences, chacunes composées de 20 images, issues de la plateforme PAVIN (séquence 1 et 2) ainsi que du simulateur ASROCAM décrit en annexe D. Les séquences initiales se composent quant à elles d'un nombre variant de 200 à 500 images. Le tableau 4.1 synthétise les résultats obtenus avec notre approche. Nous y rapportons le nombre de correspondances, la précision ainsi que le taux d'images recalées. Ce dernier est défini par le rapport du nombre d'images de la sous-séquence présentant un maximum d'appariements (pour le recalage choisi) sur le nombre total d'images formant la sous-séquence.

Sous-séquences issues de :	Précision (%)	Nombre d'appariements	Taux d'images recalées (%)
Séquence 1 PAVIN	99,8	204	100
Séquence 2 PAVIN	97,6	155	97,6
simulateur ASROCAM	97,4	237	98,3

TAB. 4.1 – Synthèse des résultats obtenus pour le recalage de sous-séquences dans une séquence intiale avec la méthode REFA3D.

Afin de comparer les performance de notre méthode, les tableaux 4.2 et 4.3 regroupent les résultats obtenus pour les approches HOG/HOF et SIFT3D.

Sous-séquences issues de :	Précision (%)	Nombre d'appariements	Taux d'images recalées (%)
Séquence 1 PAVIN	99,2	212	99,6
Séquence 2 PAVIN	96,9	178	95,3
Simulateur ASROCAM	94,8	256	92,5

TAB. 4.2 – Synthèse des résultats obtenus pour le recalage de sous-séquences dans une séquence intiale avec le HOG/HOF.

Sous-séquences issues de :	Précision (%)	Nombre d'appariements	Taux d'images recalées (%)
Séquence 1 PAVIN	98,7	284	98,1
Séquence 2 PAVIN	97,2	247	97,8
Simulateur ASROCAM	95,4	294	93,2

TAB. 4.3 – Synthèse des résultats obtenus pour le recalage de sous-séquences dans une séquence intiale avec le SIFT3D.

Il en résulte que notre approche permet d'obtenir un recalage présentant généralement la meilleure précision d'appariements et le taux d'images recalées le plus élevé. Seule la méthode SIFT3D obtient des résultats légèrement supérieurs pour la séquence 2 issue de la plateforme PAVIN. Au vu des résultats obtenus, notamment pour les séquences extraites du simulateur ASROCAM, il apparait que notre approche présente une description plus pertinente. Nous notons que ces performances sont principalement dues à nos choix d'optimisations. Le seul inconvénient réside dans la diminution du nombre de mise en correspondance.

Afin d'approfondir l'analyse des performances obtenues par notre approche et dans l'optique de son intégration dans un système de véhicules intelligents, nous étudions des recalages de sous-séquences obtenues lors d'un évitement d'obstacle. La figure 4.33 présente un échantillon de la séquence initiale et des sous-séquences utilisées, issues du simulateur ASROCAM (annexe D). Nous schématisons également les trajectoires suivies par le véhicule dans le cas idéal (trajectoire nominale sans obstacle) et lors d'un évitement d'obstacle.

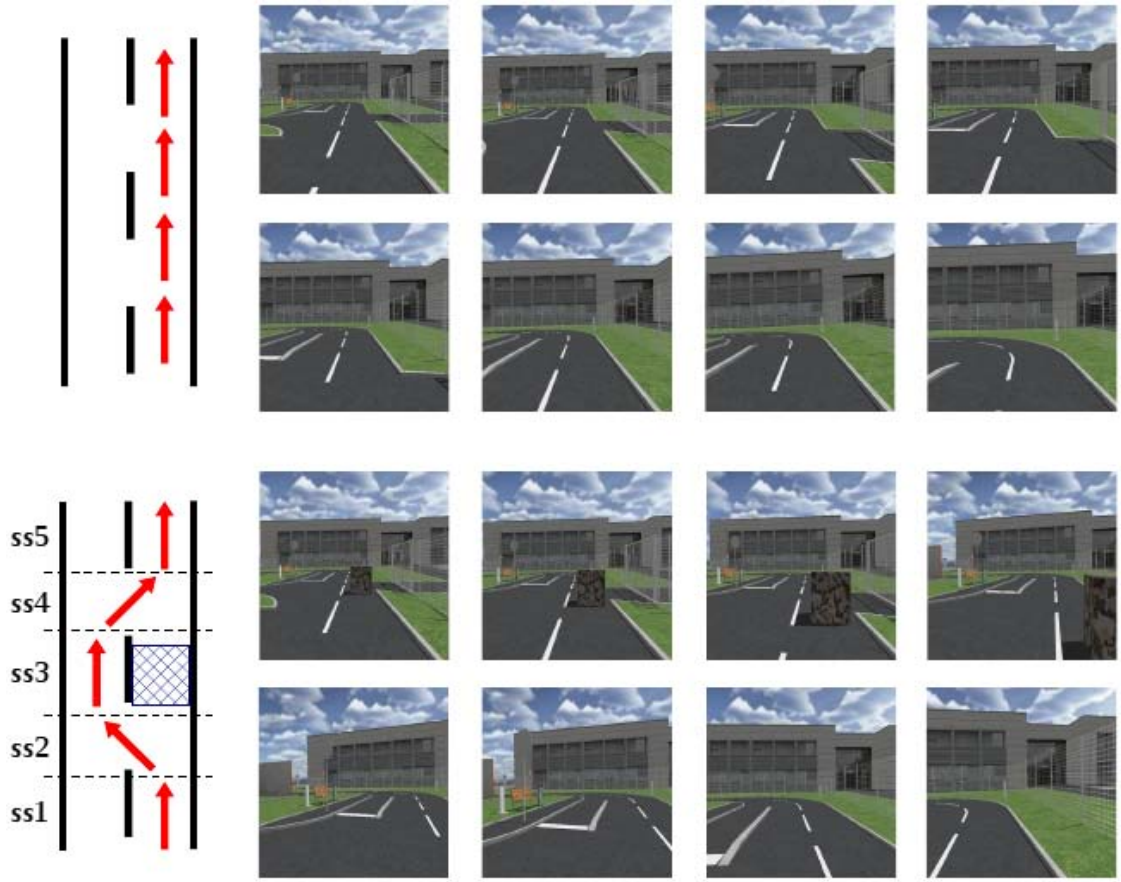


FIG. 4.33 – Echantillons de la séquence initiale et de celles présentant un évitement d'obstacle (fractionnée en cinq sous-séquence).

Le but est d'analyser les recalages obtenus au fur et à mesure de l'évitement de l'obstacle, se résumant à un changement de trajectoire du véhicule. Nous procédons à différents tests suivant cinq sous-séquences (figure 4.33 : de ss1 à ss5), permettant de représenter l'ensemble de la trajectoire suivie. Le tableau 4.4 regroupe les résultats obtenus respectivement par notre méthode, le couplage HOG/HOF et SIFT3D. Pour ces différentes approches nous étudions la précision P des appariements ainsi que le taux d'images recalées, que nous notons Tir, défini précédemment.

	REFA3D		HOG/HOF		SIFT3D	
Sous-séquences :	P (%)	Tir (%)	P (%)	Tir (%)	P (%)	Tir (%)
ss1	99,4	100	98,7	99,5	99,2	100
ss2	91,3	95,6	89,2	90,3	86,7	93,3
ss3	79,1	87,6	67,3	72,3	71,2	81,3
ss4	85,4	92,2	79,6	84,4	81,3	88,4
ss5	94,7	97,3	91,3	91,5	95,1	95,6

TAB. 4.4 – Synthèse des résultats obtenus pour le recalage de sous-séquences lors d'un évitement d'obstacle.

L'analyse de ces différents résultats démontre que notre approche obtient une précision et un taux d'images recalées généralement supérieurs à ceux des méthodes comparées. Seul SIFT3D présente, pour la sous-séquence ss5, une précision plus élevée. Notre méthode posséde également une meilleure stabilité, représentée par des diminutions plus faibles des critères d'observation, tout au long de la trajectoire. Au vu des performances obtenues par notre méthode, il serait intéressant d'envisager à plus "haut niveau" l'emploi de ces différentes données dans un processus de réalignement du véhicule sur sa trajectoire nominale. Les appariements fournis par notre approche permettrait d'estimer, image par image, l'homographie et par conséquent de calculer les différents paramètres de recalage à fournir au système de localisation.

5 Conclusion et Perspectives

Dans ce manuscrit, nous nous sommes intéressés à l'élaboration d'une méthode de description de points d'intérêt dans le but d'en extraire des appariements robustes dans le contexte d'images ou de séquences vidéo. Ce type d'analyse est couramment utilisé dans des applications telles que la reconstruction 3D, le suivi d'objets ou encore la reconnaissance de gestes. La qualité des données qui leurs sont fournies a une certaine influence sur les résultats qu'elles obtiennent. En effet, les erreurs issues de ces analyses peuvent entraîner de mauvaises estimations d'homographies (pour la reconstruction 3D) ou de déplacements (pour le suivi) par exemple. Notre objectif principal a donc été d'élaborer un système permettant d'accroître au maximum la précision des appariements.

Nous avons tout d'abord réalisé une étude bibliographique recensant un certain nombre de méthodes existantes, afin de lister leurs avantages et inconvénients. Nous avons pu établir différents constats. Concernant les détecteurs de points d'intérêt, celui proposé par Harris et Stephens [47] (détecteur de Harris) reste le plus utilisé. Le fast-hessien [13] et le Dog [73] ont quant à eux l'avantage de posséder une analyse multi-échelle, permettant notamment d'observer différents niveaux de détails. La description du voisinage du point d'intérêt s'appuie sur des histogrammes de gradients orientés (SIFT) ou des ondelettes de Haar (SURF). Ils permettent tous deux d'analyser les différents gradients constituant le voisinage du point. L'étape de mise en correspondance repose le plus souvent sur la minimisation de la distance inter-descripteur. Cette dernière permet de déterminer les couples de points avec la meilleure ressemblance.

Cette étude nous a permi de proposer une méthode d'analyse spatiale, couplant différents outils existants à un masque de description locale adaptatif. Différentes modifications et optimisations ont été apportées afin d'accroître les performances de notre approche REFA. Tout d'abord, l'extraction de points d'intérêt s'appuie sur l'utilisation du fast-hessien présentant le meilleur taux de répétabilité. Les optimisations apportées se résument à la limitation de l'espace d'échelle d'exploration et au seuillage des scores de détection (suppression des scores trop faibles). La description du voisinage des points, partie la plus innovante de notre méthode, repose sur la construction d'histogrammes de gradients orientés. Pour ce faire, nous proposons un masque adaptatif s'appuyant sur une géométrie elliptique (17 ellipses). Cette dernière permet d'obtenir une meilleure analyse de l'information présente dans le voisinage. Différents paramètres sont alors déterminés, tels que l'angle de recalage (θ), les échelles locales (σ_1 et σ_2), permettant

une adaptation du masque d'analyse propre à chaque point d'intérêt. Les gradients formant les histogrammes sont normalisés et seuillés, augmentant la robustesse aux éventuels changements de luminosité. La mise en correspondance s'appuie quant à elle sur la construction d'un arbre de décision. Cet outil diminue les temps de calculs et simplifie la recherche de minima pour les distances inter-descripteurs. Nous le couplons à un coefficient de validation, permettant d'observer la pertinence des appariements. Nous ajoutons une méthode de suppression des doublons, gérant ainsi l'unicité des mises en correspondance.

Les nombreux tests et résultats détaillés au chapitre 3 mettent en avant les performances de notre méthode pour l'analyse et la mise en correspondance spatiale. Afin de valider au mieux notre approche et dans un soucis de diversité, un grand nombre d'images et de transformations a été proposés (rotations, changements d'échelle...). L'interprétation des résultats obtenus permet d'observer pour notre méhode une précision d'appariements supérieure à celle du SIFT et du SURF. Elle obtient également un taux de mise en correspondance plus élevé, montrant ainsi la qualité des points extraits par le détecteur.

En définitive, l'utilisation de notre masque adaptatif elliptique fournit une analyse plus pertinente de l'information présente dans le voisinage. Les différents seuils et optimisations que nous avons proposés permettent quant à eux d'accroître les performances des outils existants. Le seul inconvénient observé réside dans la diminution du nombre de points appariés. En effet les différents apports et modifications ont pour conséquence une diminution du nombre de points d'intérêt et donc du nombre de mises en correspondance.

Nous avons constater au chapitre 3 que dans le cas d'analyses de séquences vidéo, les données spatiales ne sont pas suffisantes. Souhaitant répondre aux exigences des applications liées aux véhicules intelligents, l'utilisation des données temporelles est indispensable. Nous avons donc proposé une généralisation spatio-temporelle de notre méthode. Pour ce faire, nous utilisons le détecteur hes-STIP (*hessian spatio-temporal interest point*), possédant la meilleure répétabilité pour ce type d'analyse. L'optimisation que nous lui apportons concerne la limitation des échelles d'exploration (spatiale et temporelle). Le masque d'analyse est également modifié afin d'intégrer la composante temporelle dans les histogrammes. Les ellipses sont pour cela converties en ellipsoïdes et nous utilisons 5 niveaux de description (figure 4.9). L'ajout d'un recalage temporel, au recalage spatial existant, permet une exploration tridimensionnelle stable de la séquence.

Afin de valider pleinement cette généralisation spatio-temporelle, nous avons proposé plusieurs tests basés sur des séquences issues d'une part, de caméra réelle (annexe B) et d'autre part, de deux simulateurs (annexes C et D). Les premiers résultats, reposant sur des transformations synthétiques apportées à la séquence, montre que notre approche obtient généralement la meilleure précision. Nous observons également une diminution moins prononcée du nombre de points appariés, caractérisant la stabilité de la méthode au fur et mesure des transformations. Nous apparions par la suite des séquences simulant des trajectoires autour du rond-point de la plateforme PAVIN.

Notre approche reste supérieure en terme de précision et de taux d'appariements par rapport au couplage HOG/HOF et SIFT3D. Ces résultats mettent en avant la robustesse de notre approche à différentes transformations (translation, changement d'échelle temporelle) engendrées par un changement de direction. De ce fait, les derniers tests que nous étudions sont liés aux applications concernant les véhicules intelligents. En effet, le recalage de sous-séquences dans une séquence initiale permet de fournir des informations spatio-temporelles au véhicule. Ces dernières sont utilisées lors de l'estimation de la distance du véhicule par rapport à sa trajectoire nominale, ou encore de sa localisation précise relative à cette même trajectoire. Les résultats obtenus par notre approche montre un fort taux de précision et une bonne stabilité. Pour des séquences présentant un évitement d'obstacle, le véhicule change de trajectoire, entraînant une accumulation de transformations par rapport à la séquence de référence. Malgré cela, notre méthode obtient les meilleurs résultats, aussi bien en terme de précision que de taux d'images recalées.

En reprenant l'ensemble des résultats, nous pouvons conclure que notre méthode possède une détection de points d'intérêt de qualité et que les descriptions elliptiques (2D) ou ellipsoïdiques (2D+t) sont précises et pertinentes. Nous mettons également en avant la bonne stabilité vis à vis d'une part, des transformations applicables à l'image et d'autre part, de la détérioration des données.

Au vu des différents travaux présentés dans ce manuscrit, un certain nombre de perspectives peut être envisagé. Ayant démontré que notre approche obtient une précision d'appariements élevée, nous envisageons de l'intégrer dans différentes applications liées aux véhicules intelligents, principalement la localisation et la reconstruction 3D de l'environnement. En effet, la qualité des données joue un role important dans leurs performances finales, il est donc essentiel de s'appuyer sur une méthode proposant une très grande précision. Un second avantage de notre approche concerne la mise en correspondance de séquences lors d'un évitement d'obstacle. Les résultats obtenus pour le recalage de sous-séquences montrent que notre approche conserve un taux d'appariements et une précision élevés, tout au long des changements de direction. L'interprétation de ces données permettra par exemple d'estimer, image par image, l'homographie. Notre approche apportera donc des informations suplémentaires au système, pouvant être utilisées pour le réalignement du véhicule sur sa trajectoire nominale par exemple.
Une généralisation spatio-temporelle de notre méthode impose l'ajout d'une troisième composante à notre approche initiale. Par conséquent, une seconde perspective serait d'exporter cette dernière au domaine tridimensionnel afin de pouvoir observer et analyser des objets 3D. Une des applications visées concerne l'imagerie 3D (scanner, IRM, échographie, scintigraphie). Les données fournies par notre approche permettront le recalage d'images sur un atlas, dans le but de segmenter l'objet observé. Elles seront également utilisées pour de processus de suivi 3D notamment en cardiologie.
Un dernier point qui n'a pas été abordé dans ce manuscrit, mais qui nous paraît interressant de préciser, concerne les temps de calculs. A l'heure actuelle, notre approche présente une durée d'exécution plus rapide que celle du SIFT pour l'analyse

2D, et que celle du SIFT3D pour l'analyse 2D+t. Concernant les deux autres méthodes auxquels nous nous sommes comparés (SURF et HOG/HOF), notre analyse est plus coûteuse. Une dernière perspective prévue à court terme est donc d'optimiser nos algorithmes en ayant recours à des librairies accélérant les temps de calculs. Actuellement en dévellopement au laboratoire, la mise à jour de NT^2 [40][41] sera intégrée à nos algorithmes. Elle permettra entre autre de paralléliser les différentes analyses, notamment les étapes de description du voisinage.

Annexes

A. Bases d'images utilisées pour l'analyse spatiale

B. Séquences vidéo issues de la plateforme PAVIN

C. Simulations de trajectoire pour l'analyse spatio-temporelle

D. Simulations de trajectoire : ASROCAM

E. Méthode d'estimation robuste de la matrice d'homographie

F. Résultats complémentaires

A Bases d'images utilisées pour l'analyse spatiale

A.1 Bases de données :

– internet :

FIG. A.1 – Exemple d'images tests extraites d'internet (Beatles, Lena, Pig).

– base de données d'Oxford[1] :

FIG. A.2 – Exemple d'images tests extraites de la base de données d'Oxford (Graffiti, Boat, Leuven, Wall, Manoir, Trees).

[1]http ://www.robots.ox.ac.uk/ vgg/data/data-aff.html

– séquences d'acquisition provenant de la plateforme PAVIN.

FIG. A.3 – Exemple d'images tests extraites d'une séquence d'acquisition (plateforme PAVIN).

A.2 Types de transformations étudiées

Ces différentes images peuvent être regroupées par type de transformation, que ce soit des modifications synthétiques ou des changements dû aux déplacements de la caméra (transformations réelles).

– rotation :

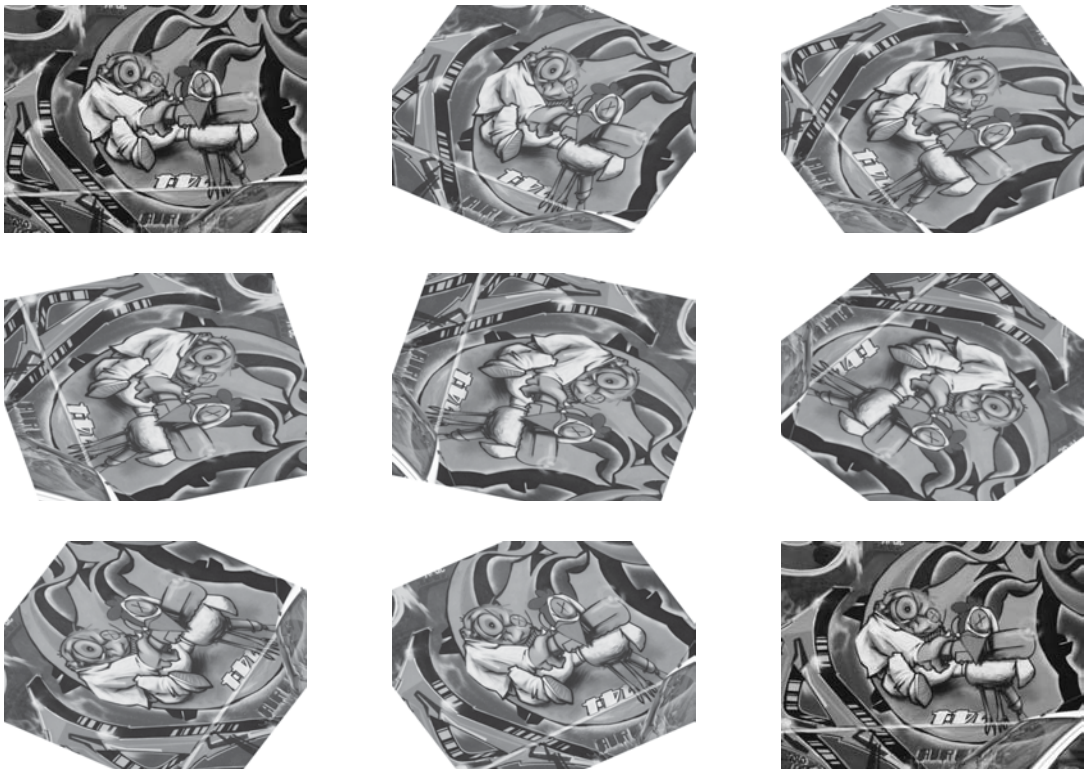


FIG. A.4 – Transformations de type rotation (appliquées aux images Graffiti).

– changement d'échelle :

FIG. A.5 – Transformations de type changement d'échelle (appliquées aux images Beatles et Lena).

– étirements :

FIG. A.6 – Transformations de type étirement (appliquées aux images Pig et Graffiti).

– couplage changement d'échelle et rotation :

FIG. A.7 – Transformations de type changement d'échelle et rotation (appliquées aux images Boat).

– changement de luminosité :

FIG. A.8 – Transformations de type changement de luminosité (appliquées aux images Leuven).

– changement de point de vue :

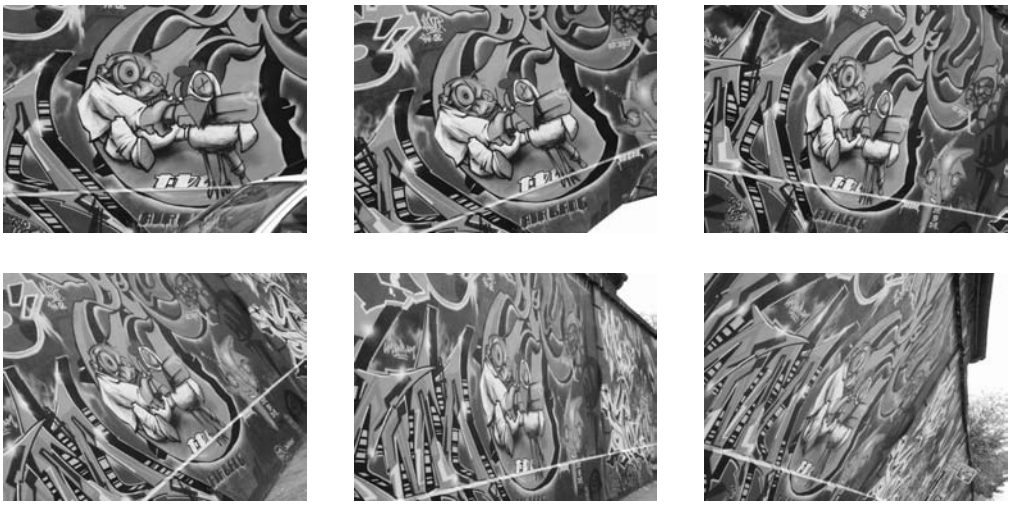


FIG. A.9 – Transformations de type changement de point de vue (appliquées aux images Graffiti).

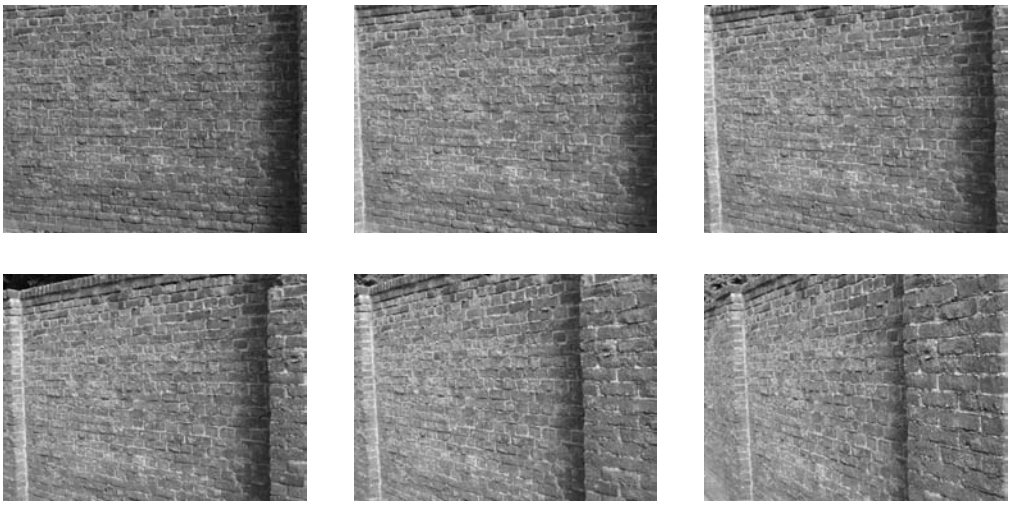


FIG. A.10 – Transformations de type changement de point de vue (appliquées aux images Wall).

– modification de compression JPEG :

FIG. A.11 – Transformations de type modification de compression JPEG (appliquées aux images Manoir).

Pour finir, nous précisons que ces différentes figures ne représentent qu'une partie de notre base de test. En effet afin d'obtenir le plus de résultats possibles, chaque type de transformations s'appuie sur un nombre d'images plus conséquent.

B Séquences vidéo issues de la plateforme PAVIN

Issue d'un projet conjointement initié en 2005 par le CNRS et le Lasmea, la plate-forme PAVIN (Plate-forme d'Auvergne pour Véhicules INtelligents) est un environement de près de 5000m^2 constitué de 317 mètres de voirie, d'un rond-point, de carrefours et de façades représentant ainsi un milieu urbain. Cette plate-forme permet au LASMEA d'approfondir ses recherches sur les véhicules intelligents, la localisation ou encore la reconstruction 3D de l'environnement. La figure B.1 présente une photographie de cette plate-forme.

FIG. B.1 – Photographie de la plate-forme PAVIN (octobre 2009).

Concernant nos travaux de recherche, ayant pour perspective d'intégrer la méthode REFA3D aux véhicules intelligents, une partie de nos tests s'appuie sur des séquences issues de cete plateforme. La figure B.2 présente les trois trajectoires retenues pour la validation de notre approche.

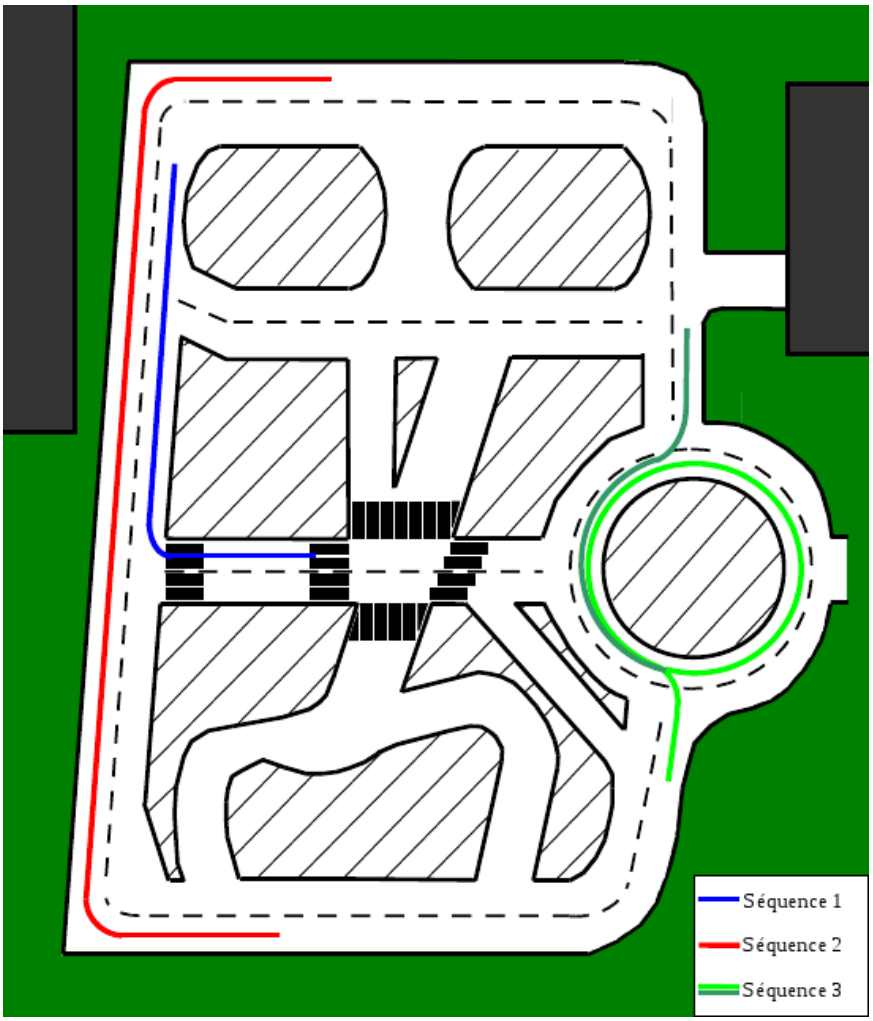


FIG. B.2 – Schéma de la plate-forme PAVIN et trajectoires tests qui en sont extraites.

La première séquence est constituée d'un trajet à vitesse constante et profite d'un faible changement de luminosité tout au long du parcours. La deuxième séquence se divise en trois parties : une phase à vitesse réduite suivie d'une accélération tout au long de la ligne droite pour ensuite se terminer par un nouveau ralentissement. Cette séquence intégre également un changement de luminosité entraînant l'apparition d'ombres pouvant perturber la description locale. La dernière séquence s'illustre quant à elle par l'utilisation d'un rond-point créant ainsi un jeu de données similaires lors du second passage du véhicule. Ce trajet possède également des changements d'intensité lumineuse ainsi que des modifications d'échelles temporelles.

Les figures B.3 et B.4 présentent un échantillon des images issues respectivement des séquences 1 et 2 (schématisées en figure B.2). La figure B.5 regroupe quant à elle une partie des images du trajet effectué par le véhicule autour du rond-point (séquence 3 de la figure B.2).

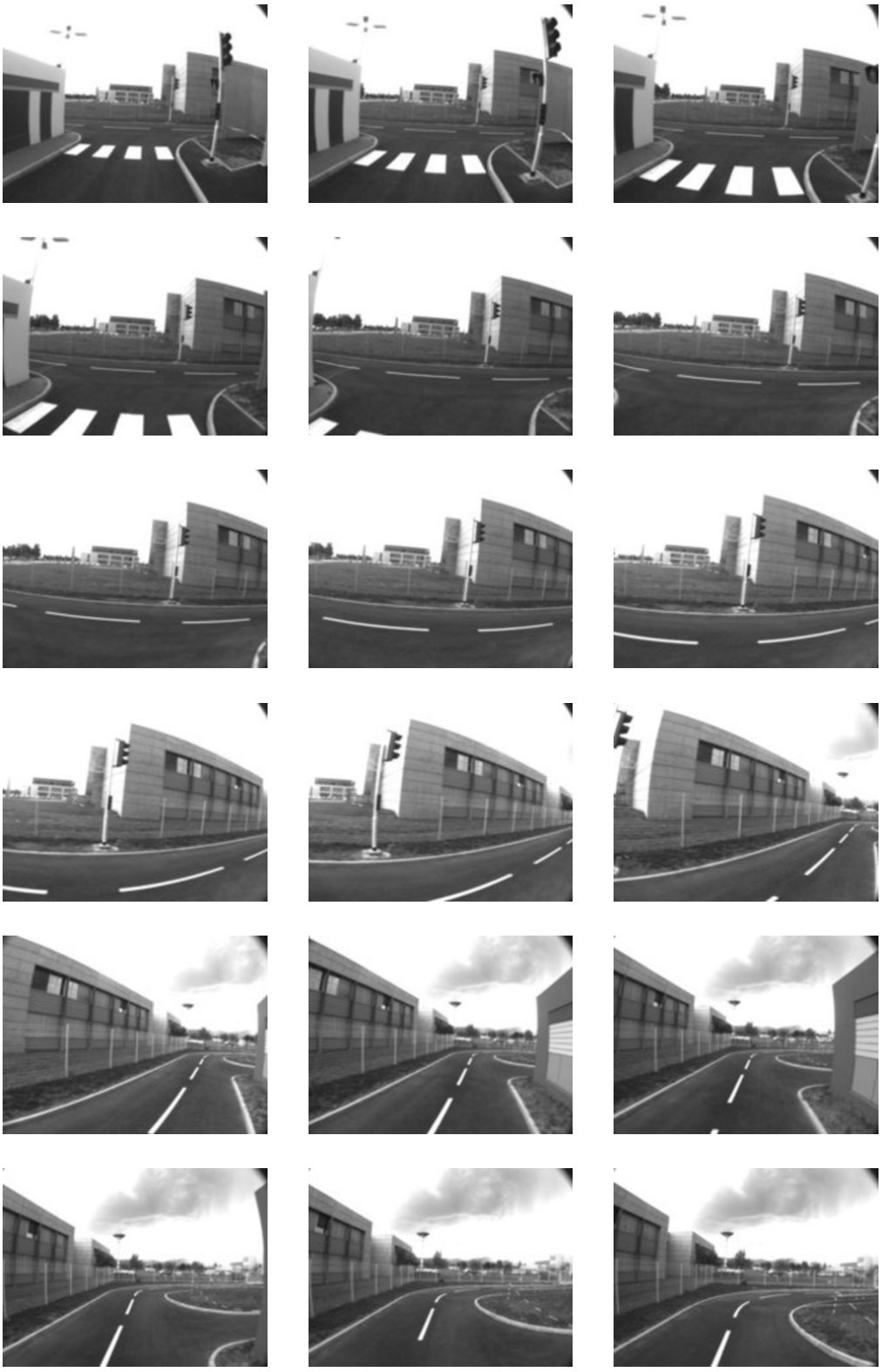


FIG. B.3 – Echantillon d'images de la séquence 1 issue de la plate-forme PAVIN.

FIG. B.4 – Echantillon d'images de la séquence 2 issue de la plateforme PAVIN.

FIG. B.5 – Echantillon d'images de la séquence 3 issue de la plateforme PAVIN.

C Simulateur de trajectoire pour l'analyse spatio-temporelle

Afin de valider la généralisation spartio-temporelle de notre approche nous utilisons l'application suivante permettant de créer un environnement virtuel et de similer une trajectoire. La figure C.1 schématise l'environnement créé ainsi que le parcours effectué par notre véhicule. Ce dernier possède deux caméras permettant d'extraire une vue avant et une vue arrière.

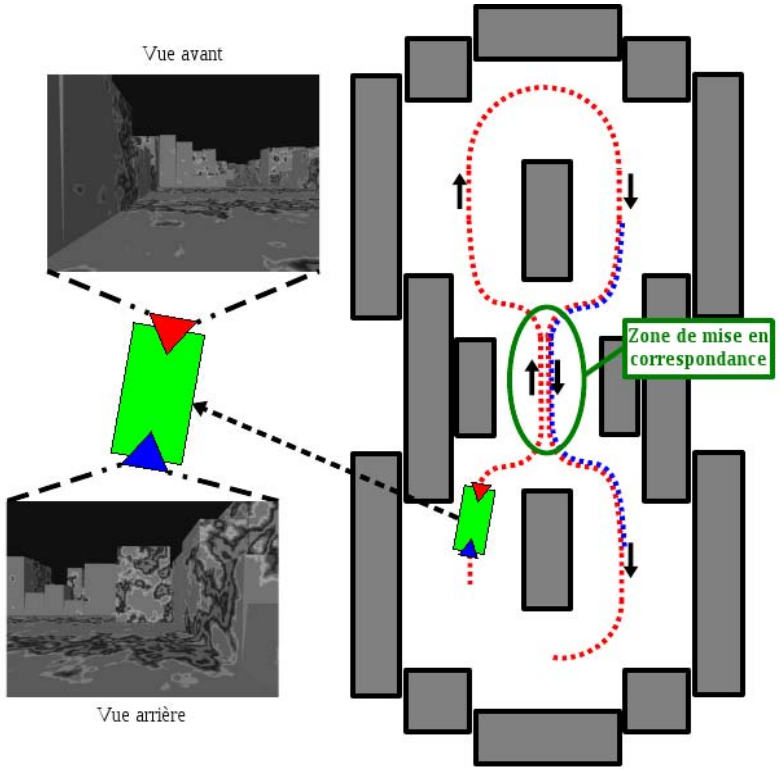


FIG. C.1 – Création d'un environnement virtuel et d'une trajectoire test. L'utilisation de deux caméras permet une observation double de la scène.

Les figures C.2 et C.3 présentent un échantillon des images extraites des deux caméras tout au long du trajet.

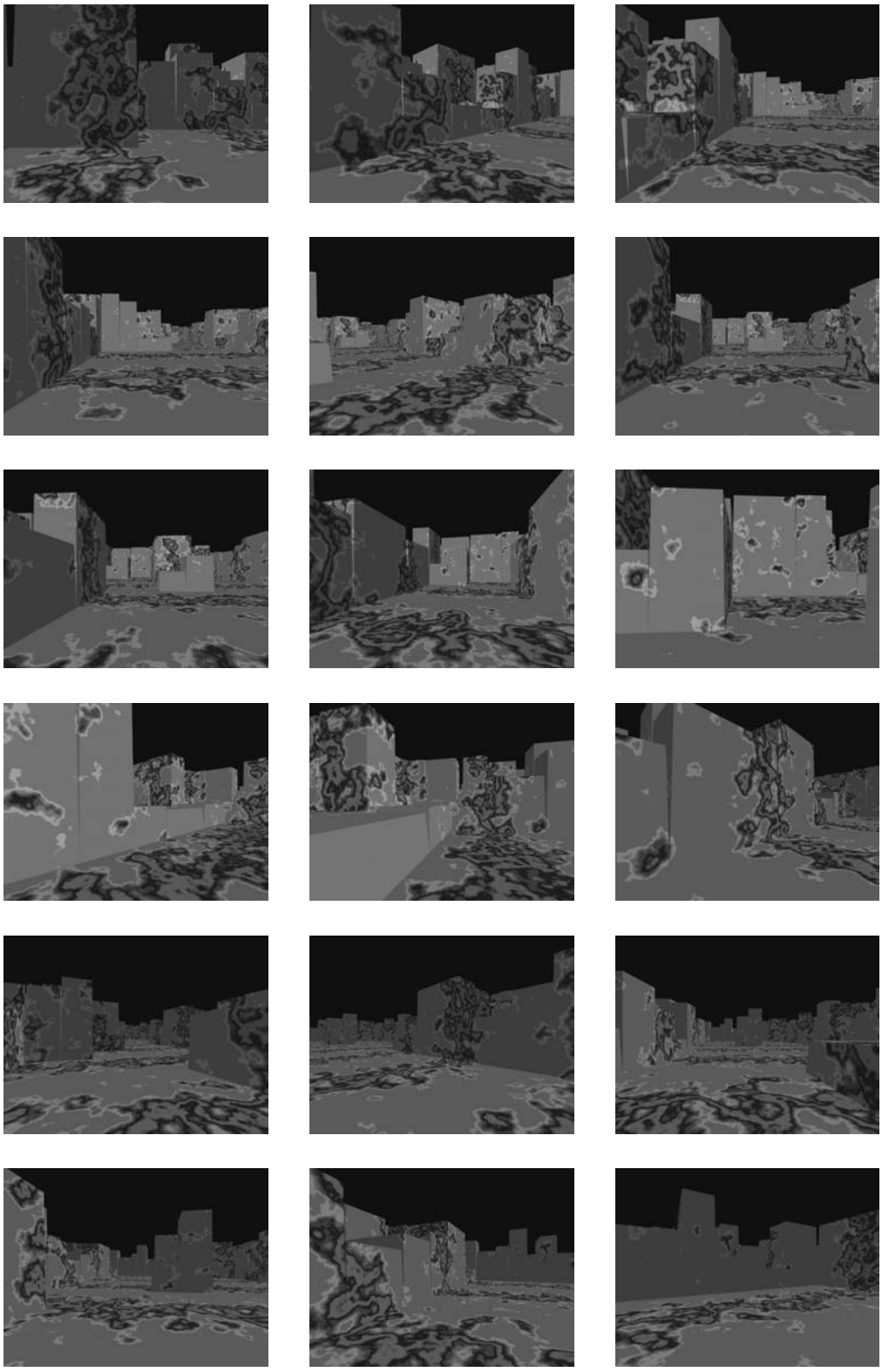


FIG. C.2 – Echantillon de la séquence présentant la vue avant du véhicule tout au long du trajet.

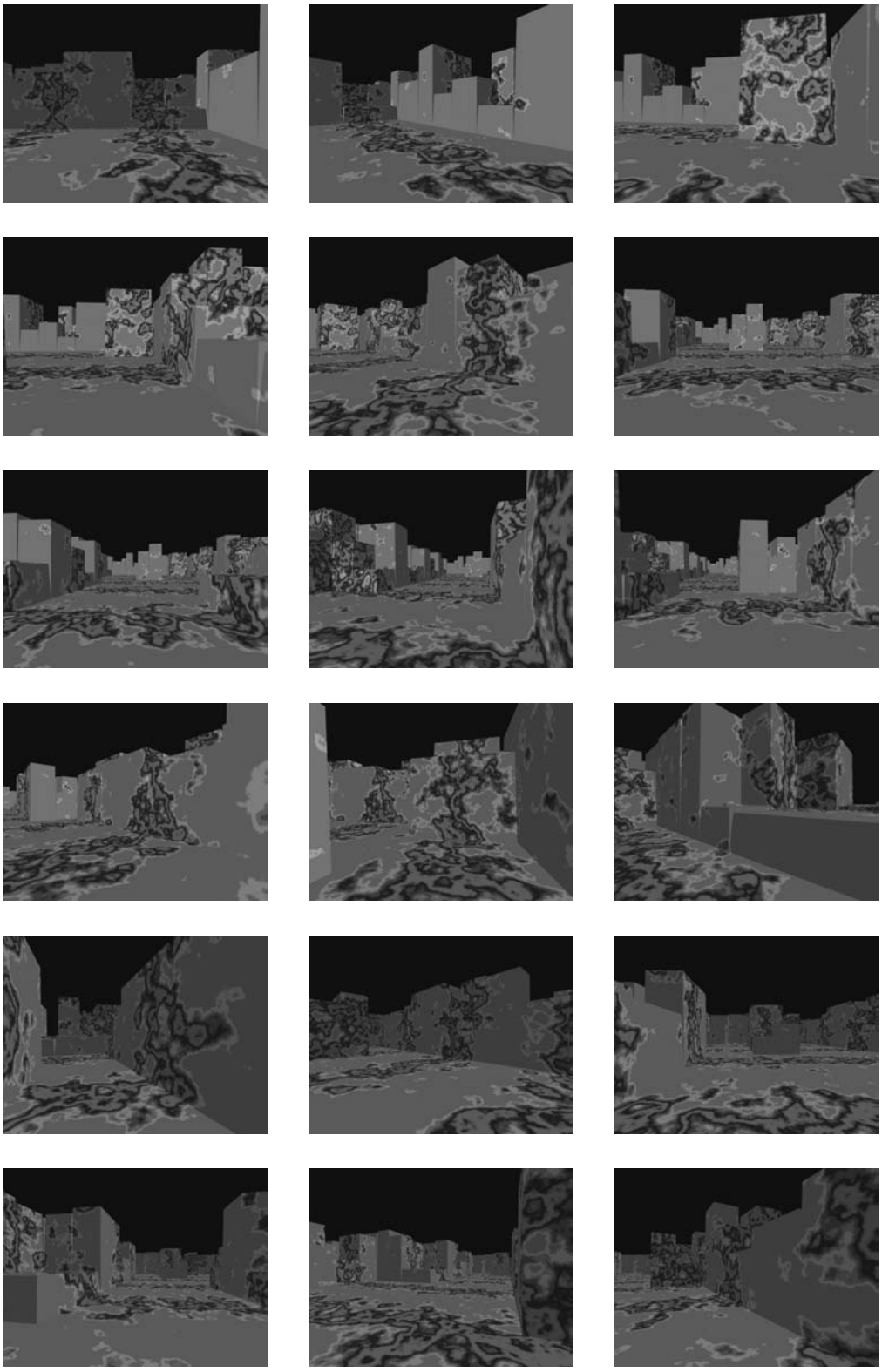


FIG. C.3 – Echantillon de la séquence présentant la vue arrière du véhicule tout au long du trajet.

D Simulations de trajectoire : ASROCAM

Développé par Florent Malartre [75] et Pierre Delmas [33] au sein d'une collaboration LASMEA/CEMAGREF, le simulateur ASROCAM permet de créer des simulations de trajectoire, à l'aide d'un joystick, dans un environnement virtuel. Ce dernier correspond à la représentation de la plateforme PAVIN (figure D.1) dans laquelle trois trajectoires sont mises en avant autour du rond-point : "intérieure", "centrée" et "extérieure".

FIG. D.1 – Simulation de 3 trajectoires ("intérieure", "centrée" et "extérieure") autour du rond-point de la plateforme PAVIN

Les figures D.2, D.3 et D.4 présentent un échantillon des images extraites des trois trajectoires.

FIG. D.2 – Echantillon de la séquence présentant la trajectoire "intérieure" du véhicule autour du rond-point.

FIG. D.3 – Echantillon de la séquence présentant la trajectoire "centrée" du véhicule autour du rond-point.

FIG. D.4 – Echantillon de la séquence présentant la trajectoire "extérieure“ du véhicule autour du rond-point.

Nous utilisons également deux séquences composées d'une part, d'une trajectoire nominale du véhicule (figure D.5) et d'autre part, s'une trajectoire présentant l'évitement d'un obstacle (figure D.6).

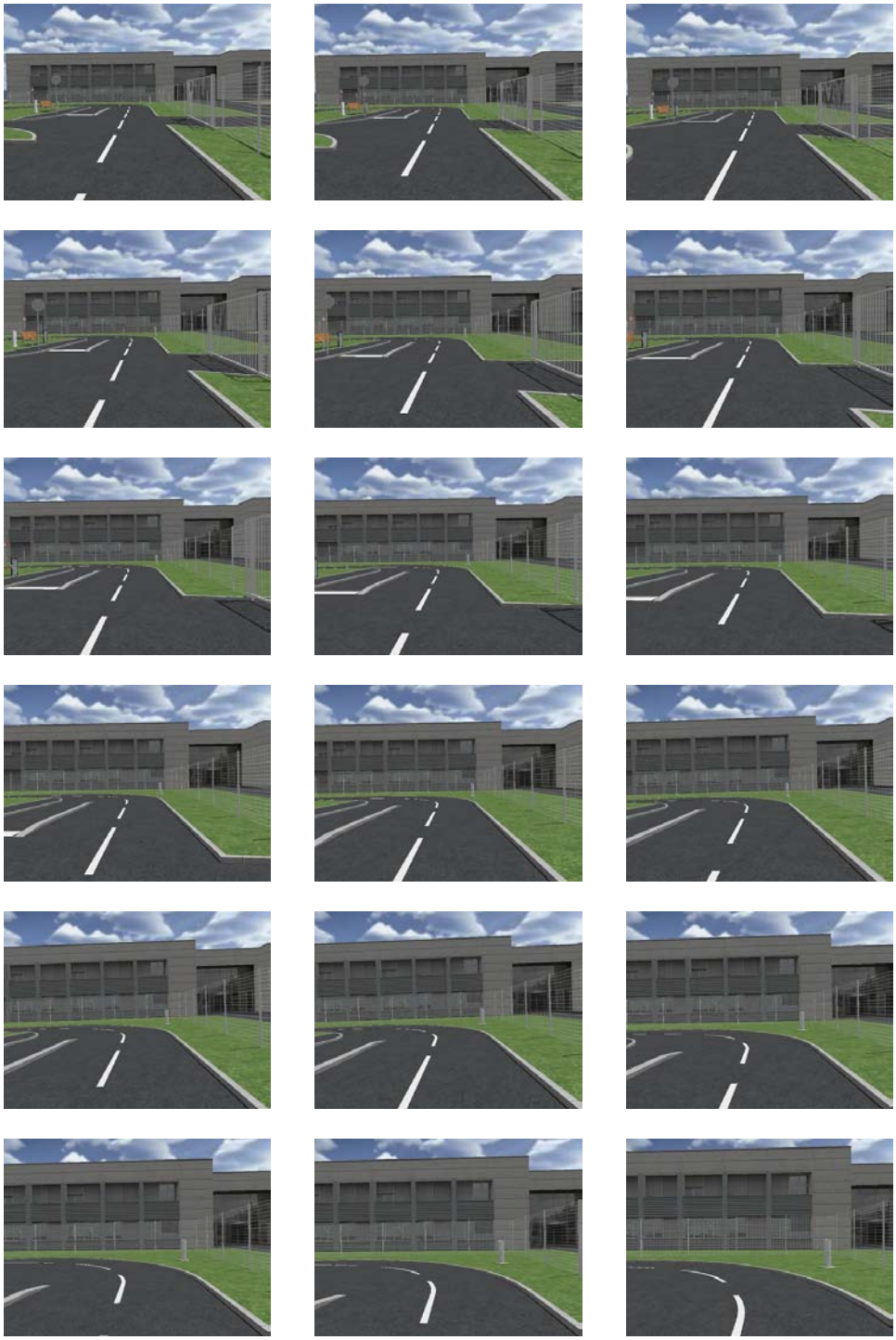


FIG. D.5 – Echantillon de la séquence présentant la trajectoire nominale du véhicule.

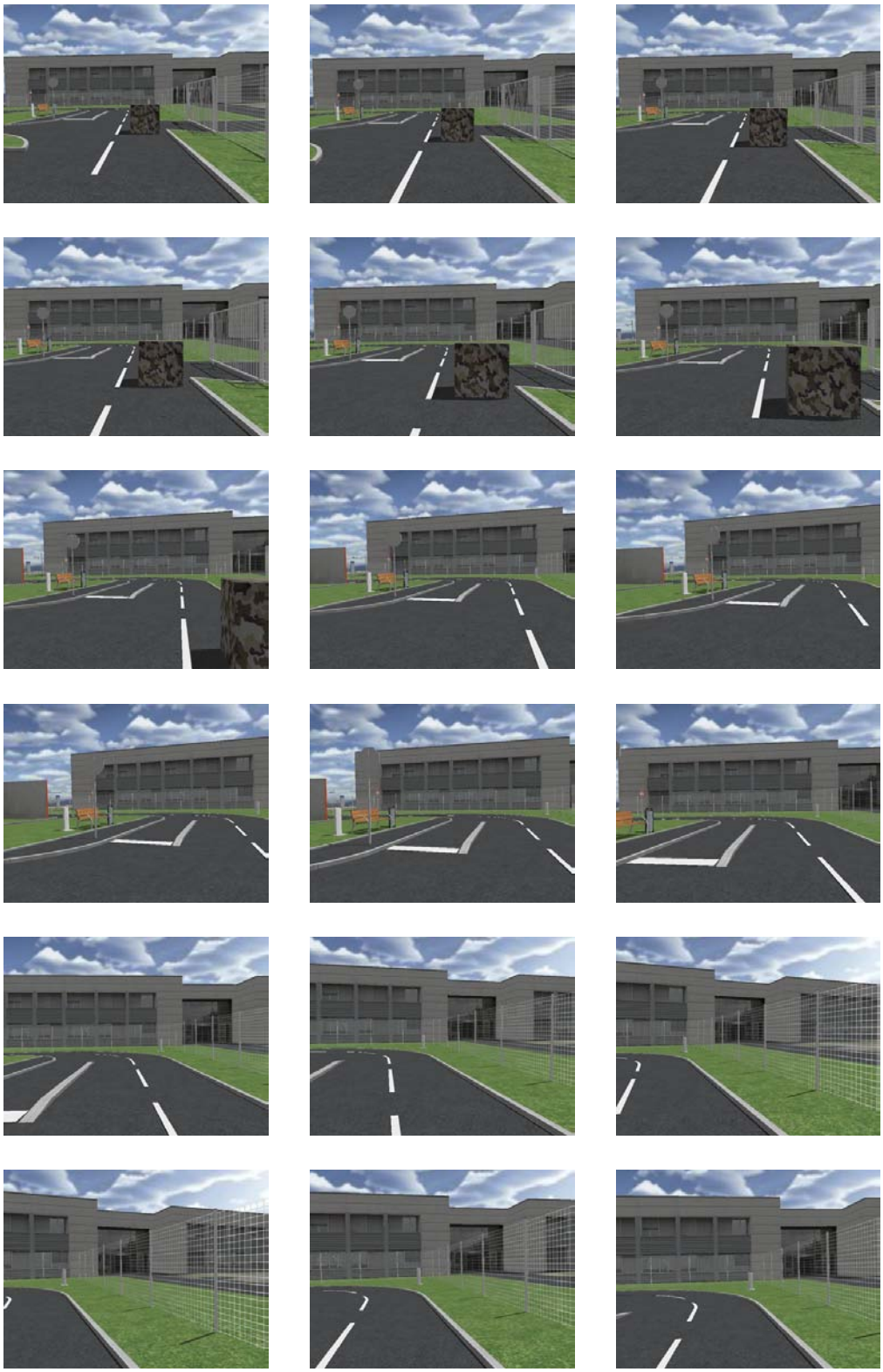


FIG. D.6 – Echantillon de la séquence présentant la trajectoire avec évitement de l'obstacle.

E Méthode d'estimation robuste de la matrice d'homographie

Les processus d'estimation de matrice d'homographie sont nombreux (Hartley et Zisserman [48], Benhimane et Malis [16] par exemple) et sont couramment utilisés en vision par ordinateur. Nous proposons dans cette annexe une liste non-exhaustive (basée sur les travaux de Malis et Marchand [76]) des méthodes existantes. Souhaitant se limiter aux algorithmes d'estimation robuste, nous différencions deux approches. La première résout simultanément le problème des *outliers* et l'estimation des paramètres de la matrice (LMS [95][96], LTS [96], méthodes de M-estimateurs [14][54]). La seconde approche consiste à séparer la détection des *outliers* et les étapes d'estimation en intégrant une notion de vote (transformée de Hough [52], méthode RANSAC [42]).

E.1 Approches par résolution simultanée :

Le principe de ces approches consiste à minimiser une fonction de coût de façon itérative afin d'estimer les paramètres de la matrice d'homographie.

- ***Least Median of Squares*** (LMS) : Introduite par Rousseeuw et Leroy [95][96], cette approche correspond à une reformulation du problème de régression linéaire standard. La fonction de coût à minimiser est définie par :

$$c(\mathbf{p}) = \text{median}(r_1^2(\mathbf{p}), r_2^2(\mathbf{p}), ..., r_n^2(\mathbf{p})), \tag{E.1}$$

  où $r_i(\mathbf{p})$ correspond aux réponses (communément appelées résidus) obtenues avec l'ensemble des données (dans notre cas les couples de points d'intérêt) sur la matrice d'homographie définie par les paramètres $\mathbf{p}$. Ne prenant en compte que la moitié des réponses ($n = N/2$ avec N le nombre total de données), cette approche est par conséquent robuste à 50% de correspondances erronées.

- ***Least Trimmed Squares*** (LTS) : Dans une optique d'accélérer la vitesse de convergence de la méthode précédemment citée, Rousseeuw et Leroy [96] proposent de rendre adaptatif l'estimation de la matrice d'homographie. Cette méthode cherche à minimiser la somme des carrés des k premiers résidus rangés par ordre

croissant :

$$c(\mathbf{p}) = \sum_{i=1}^{k} (r_i^2(\mathbf{p})). \tag{E.2}$$

Il est par conséquent possible d'estimer un pourcentage d'*outliers* présents dans les correspondances.

– **M-estimateurs** : Ces approches cherchent à minimiser l'influence des réponses présentant les plus grandes valeurs en les pondérant par une fonction ρ. La fonction de coût étudiée est par conséquent modifiée et devient :

$$c(\mathbf{p}) = \sum_{i=1}^{N} \rho(r_i(\mathbf{p})). \tag{E.3}$$

Nous choissisons de retenir deux fonctions de pondération, la première introduite par Huber [54] est définie par :

$$\rho(r_i(\mathbf{p})) = \begin{cases} \frac{1}{2} r_i^2(\mathbf{p}) & \text{si } r_i^2(\mathbf{p}) \leqslant 1,345\sigma \\ 1,345\sigma(|r_i(\mathbf{p})| - \frac{1,345\sigma}{2}) & \text{si } r_i^2(\mathbf{p}) > 1,345\sigma \end{cases}, \tag{E.4}$$

où $\sigma = 1,48 \times \text{median}(|\mathbf{r} - \text{median}(\mathbf{r})|)$ correspond à une estimation robuste de l'écart-type et $\mathbf{r}$ regroupe l'ensemble des résidus ordonnés par ordre croissant. La seconde fonction de pondération, proposée par Beaton et Tukey [14] se base sur :

$$\rho(r_i(\mathbf{p})) = \begin{cases} \frac{(4,6851\sigma)^2}{6}(1 - (1 - (\frac{r_i(\mathbf{p})}{i})^2)^3) & \text{si } r_i^2(\mathbf{p}) \leqslant 4,6851\sigma \\ \frac{(4,6851\sigma)^2}{6} & \text{si } r_i^2(\mathbf{p}) > 4,6851\sigma \end{cases}. \tag{E.5}$$

E.2 Approches basées sur une méthode de vote :

Le principe de ces approches vise à réaliser un certain nombre d'estimations de la matrice d'homographie (retournant chacune un vote) pour que l'un des ensembles de correspondances ne contienne que des *inliers* (vote le plus élevé). La matrice résultante ne dépend d'aucun *outlier*, elle est par conséquent optimale.

– **Transformée de Hough** : Décrite en 1959 par Hough [52] puis modifiée en 1988 par Illingworth et Kittler [56], cette méthode est parfaitement adaptée aux problèmes possédant un nombre important de données par rapport au nombre de paramètres à estimer. Dans le contexte d'une estimation d'homographie, cette méthode devient très coûteuse du fait des huit ou neuf paramètres mis en jeu. En effet, le principe est de travailler dans l'espace des paramètres en discrétisant ces

derniers. Nous observons que la création, l'accumulation de votes et l'estimation dans un espace $\mathbb{R}^8$ ou $\mathbb{R}^9$ entrainent une forte augmentation des temps de calculs.

– **RANSAC** : La méthode RANSAC (*RANdom SAmple Consensus*) introduite par Fischler et Bolles [42] est une approche probabiliste permettant notamment de diminuer les temps de calculs de l'estimation. Elle repose sur une fonction de coût non-linéaire annulant les résidus avec la plus grande valeur (les *outliers* si l'estimation est correcte). Elle est définie par :

$$c(\mathbf{p}) = \sum_{i=1}^{N} \rho(r_i(\mathbf{p})), \tag{E.6}$$

dont la fonction de pondération est donnée par :

$$\rho(r_i(\mathbf{p})) = \begin{cases} 0 & \text{si } r_i^2(\mathbf{p}) \leqslant 2,5\sigma^2 \\ \\ 1 & \text{si } r_i^2(\mathbf{p}) > 2,5\sigma^2 \end{cases} . \tag{E.7}$$

Le principe de cette méthode est de réitérer le calcul de cette fonction de coût, estimée à partir de n correspondances tirées aléatoirement sur les N existantes, jusqu'à obtenir un tirage contenant un nombre d'*outliers* inférieur à un seuil. Une optimisation permet en se basant sur la probabilité Q qu'aucun *outlier* ne soit présent dans le tirage, de déterminer le nombre de tirages aléatoires m nécessaires :

$$m = \frac{\log(1-Q)}{\log(1-(1-\text{pourcentage d'}\textit{outliers})^n)}. \tag{E.8}$$

Dans le cas d'une estimation de matrice d'homographie, le nombre de correspondances nécessaire n est fixé à 4. Il est par conséquent possible d'établir le tableau E.2 présentant l'influence du pourcentage d'outliers et de la précision souhaitée sur le nombre de tirages nécessaires.

	Taux d'*outliers*							
	10%	20%	30%	40%	50%	60%	70%	80%
$Q = 95\%$	3	6	11	22	47	115	368	1870
$Q = 99\%$	4	9	17	33	71	178	566	2875

TAB. E.1 – Nombre de tirages aléatoires m nécessaires, dépendant du pourcentage d'*outliers* et de la précision Q.

Une généralisation de cette approche a été proposée par Rabin et al. [92] (nommée MAC-RANSAC) permettant notamment la détection de transformations multiples entre différentes images.

F Résultats complémentaires

F.1 Ajout de tests pour la validation 2D de notre méthode

Nous avons proposé dans notre manuscrit un ensemble de tests, comparant notre approche aux versions non modifiées des méthodes SIFT et SURF. L'ensemble des paramètres utilisés pour ces dernières correspondent à ceux fournit par leurs auteurs. Afin de valider pleinement notre méthode et dans un soucis d'égalité en terme de possibilités d'appariements, nous proposons de seuiller les scores de détection du SIFT et du SURF. Ce procédé nous permet d'obtenir initialement un nombre de points détectés identique pour les trois approches comparées.

F.1.1 Transformations synthétiques

– Transformations de type **changements d'échelles isotropes**.

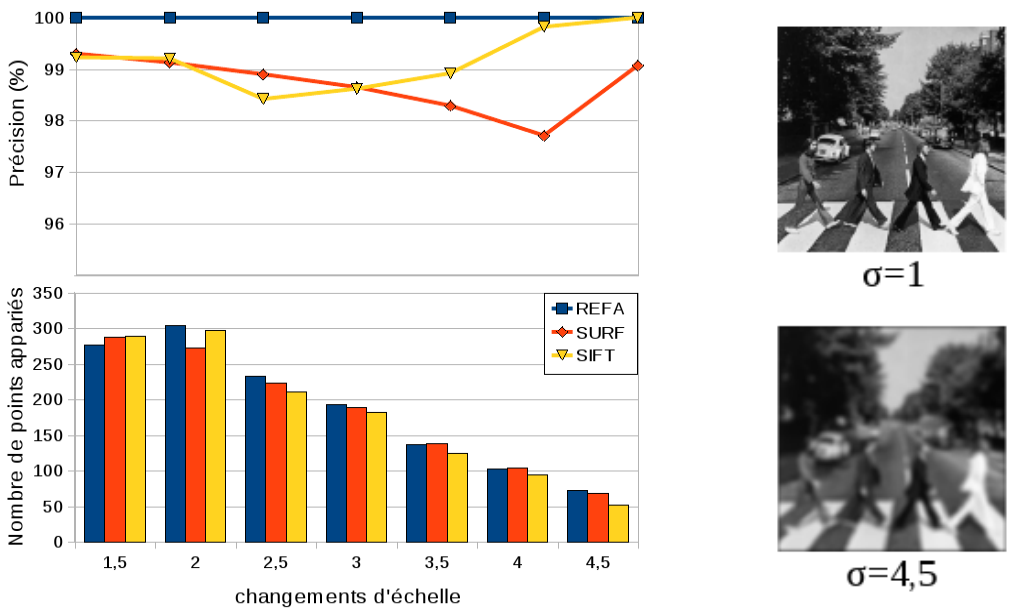


FIG. F.1 – Précision et nombre de points appariés pour des changements d'échelle sur les images Beatles.

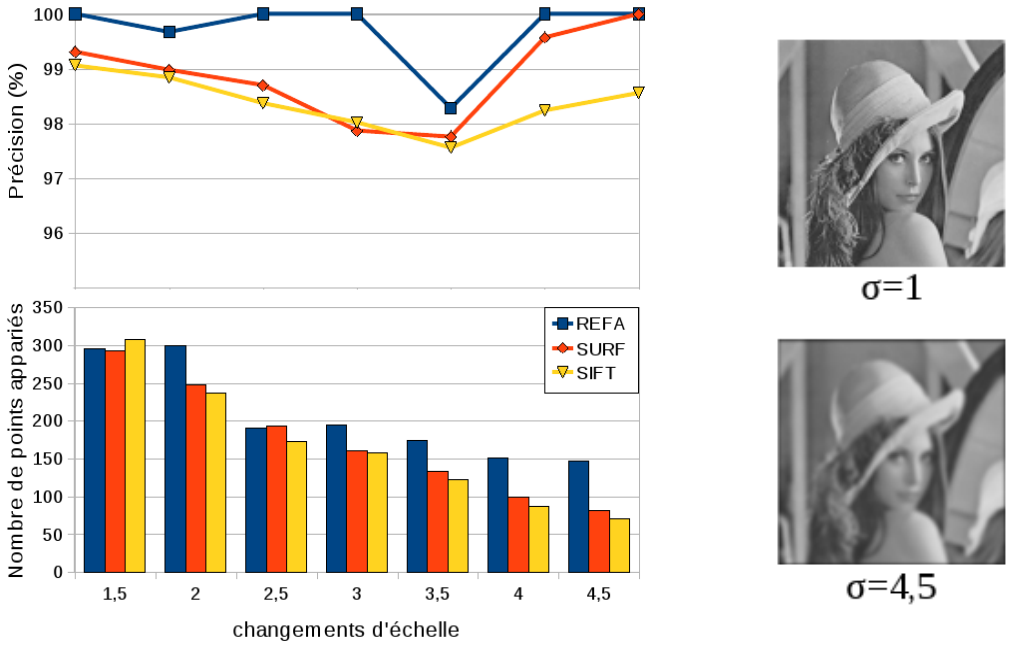


FIG. F.2 – Précision et nombre de points appariés pour des changements d'échelle sur les images Lena.

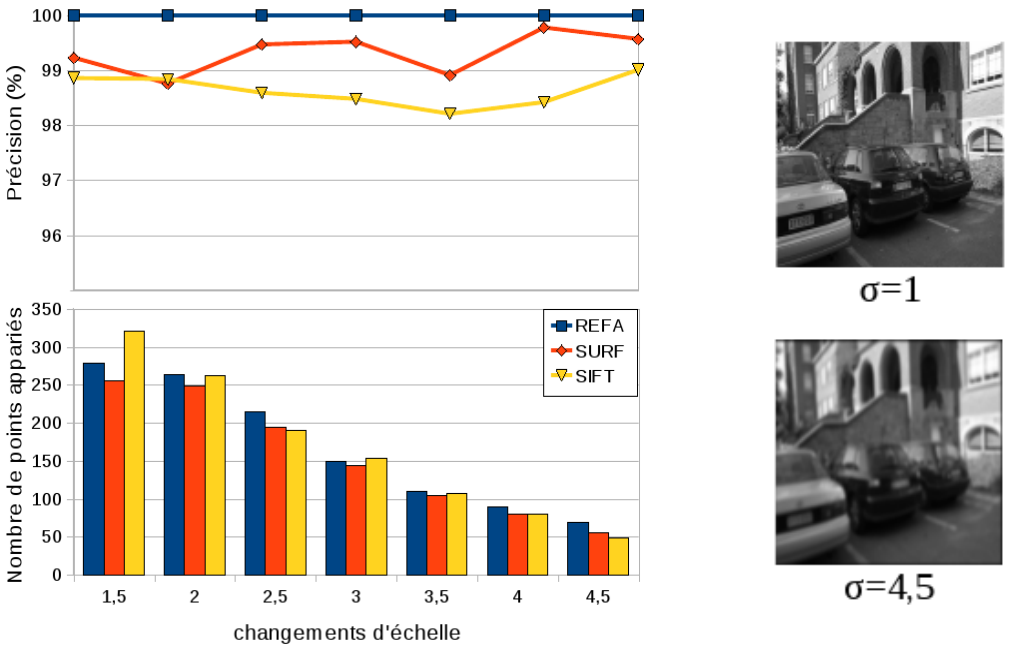


FIG. F.3 – Précision et nombre de points appariés pour des changements d'échelle sur les images Leuven.

– Transformations de type **étirements unidirectionnels**.

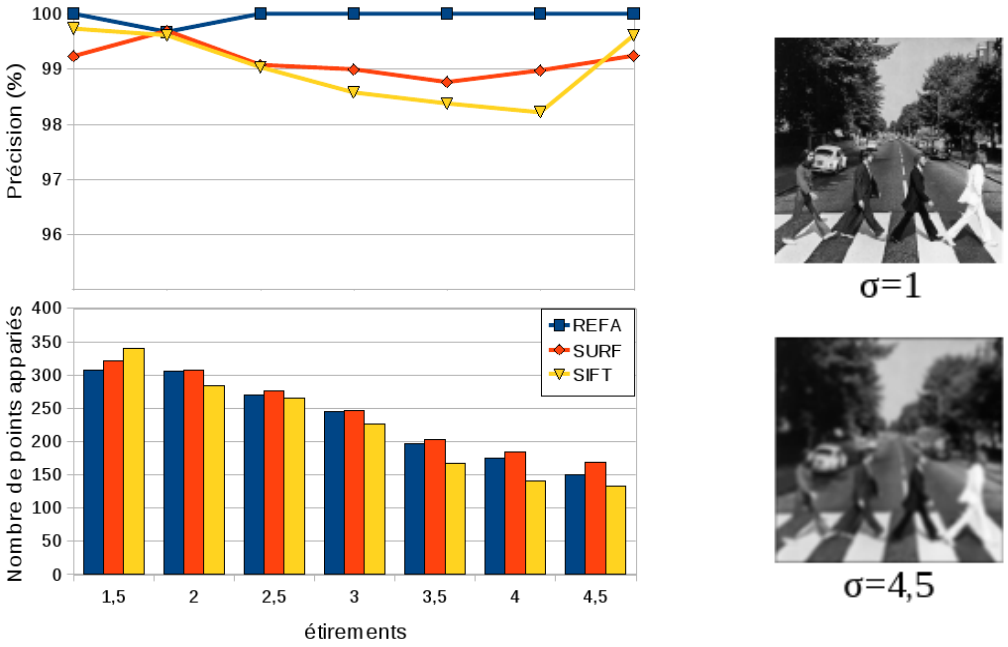


FIG. F.4 – Précision et nombre de points appariés pour des étirements unidirectionnels sur les images Beatles.

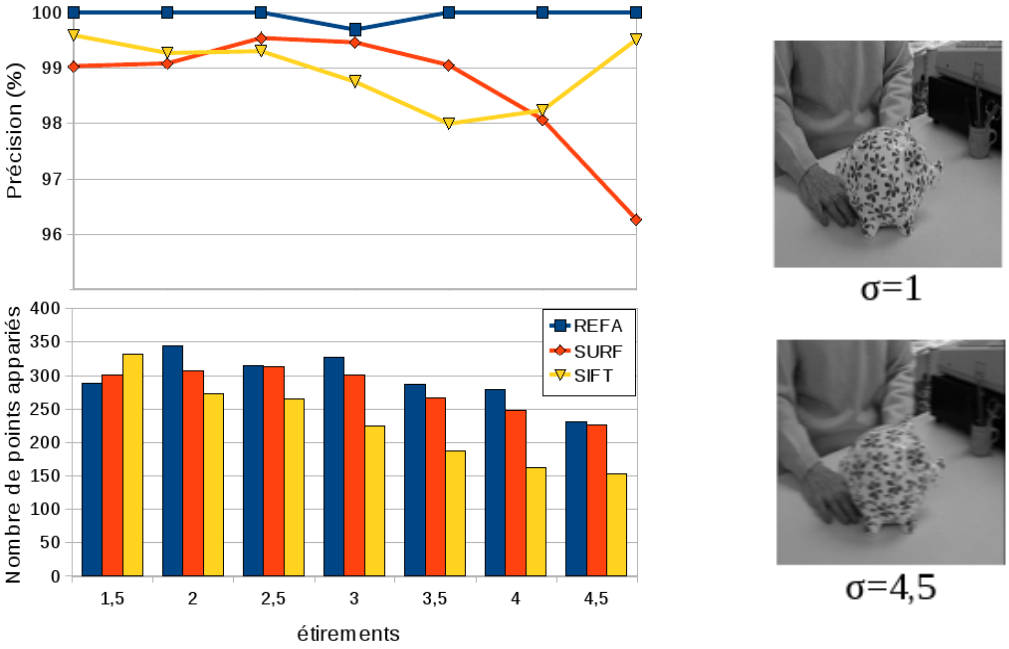


FIG. F.5 – Précision et nombre de points appariés pour des étirements unidirectionnels sur les images Pig.

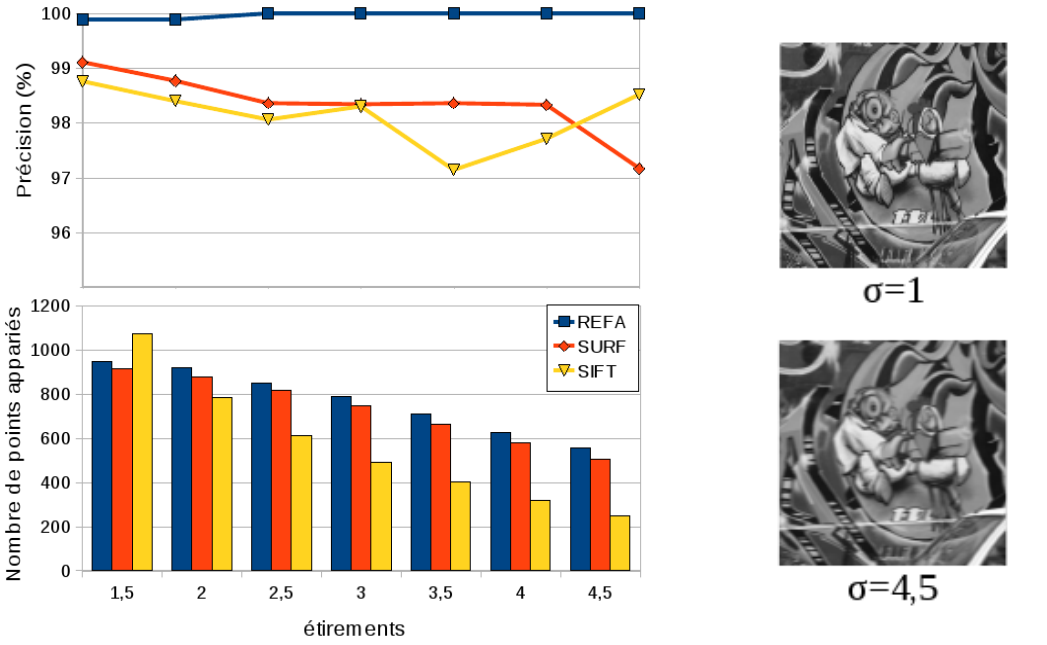


FIG. F.6 – Précision et nombre de points appariés pour des étirements unidirectionnels sur les images Graffiti.

– Transformations de type **rotations**.

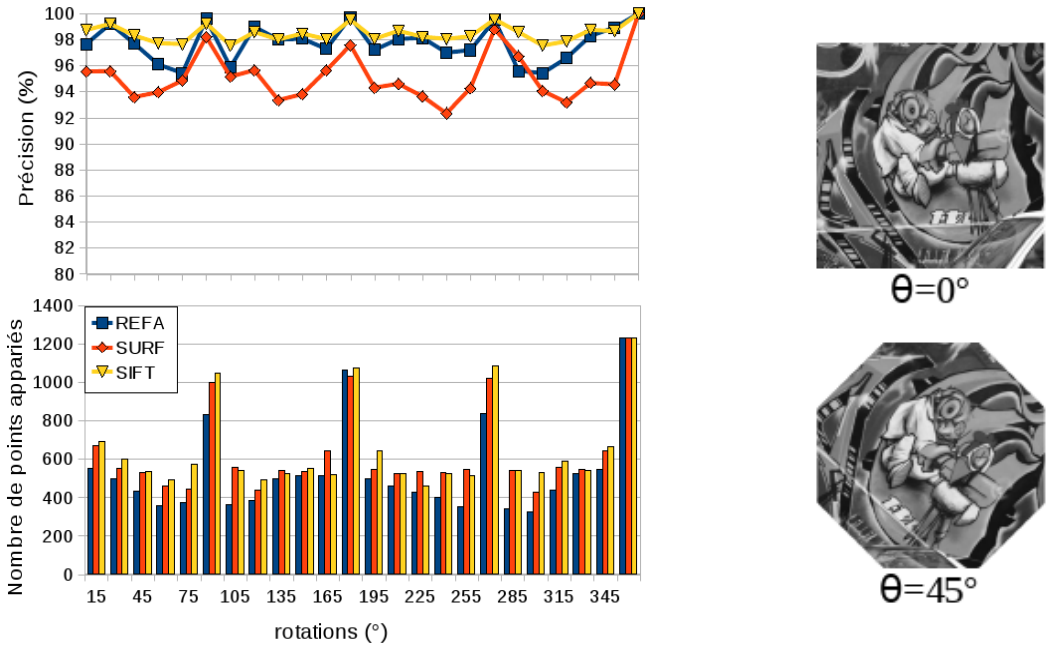


FIG. F.7 – Précision et nombre de points appariés pour des transformations synthétiques de type rotation (images Graffiti).

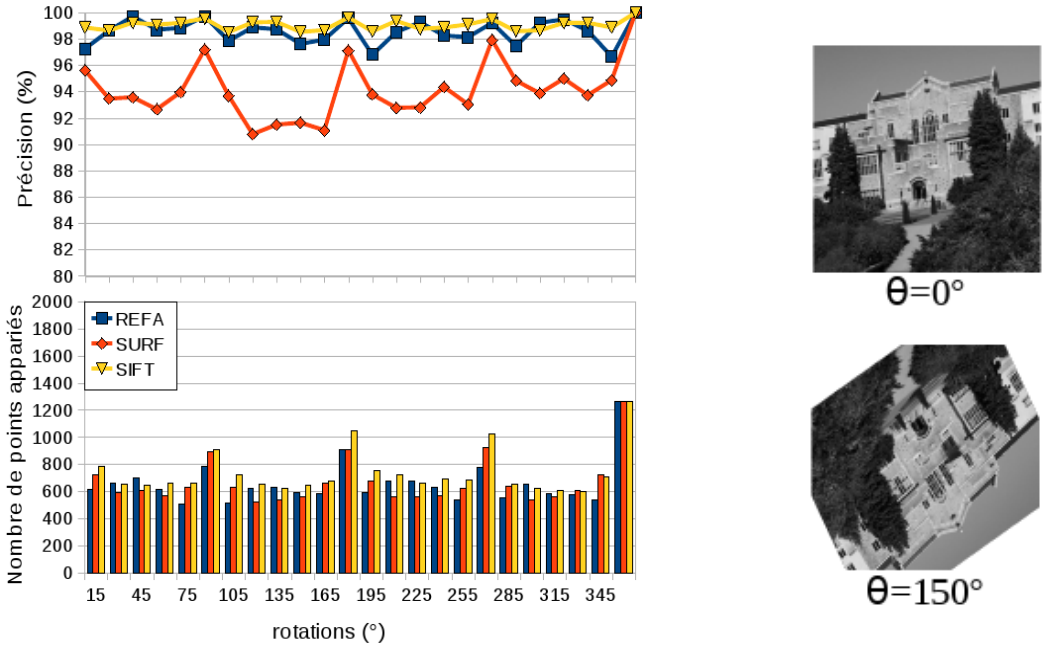


FIG. F.8 – Précision et nombre de points appariés pour des transformations synthétiques de type rotation (images Ubc).

– Transformations composées de **changements d'échelle + rotations**.

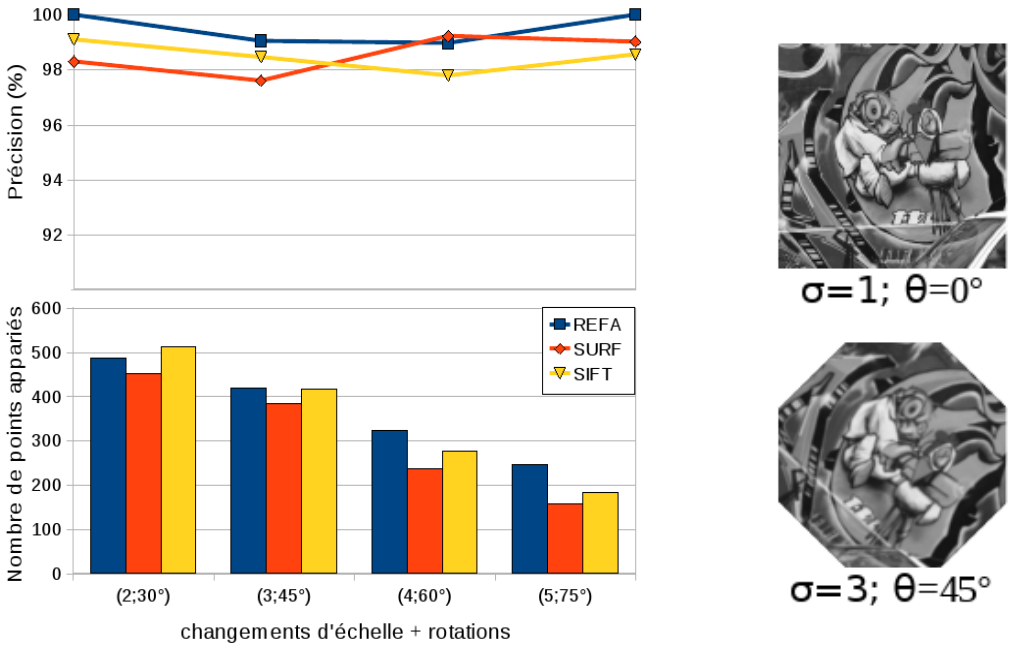


FIG. F.9 – Précision et nombre de points appariés pour un couplage changements d'échelle/rotations (Graffiti).

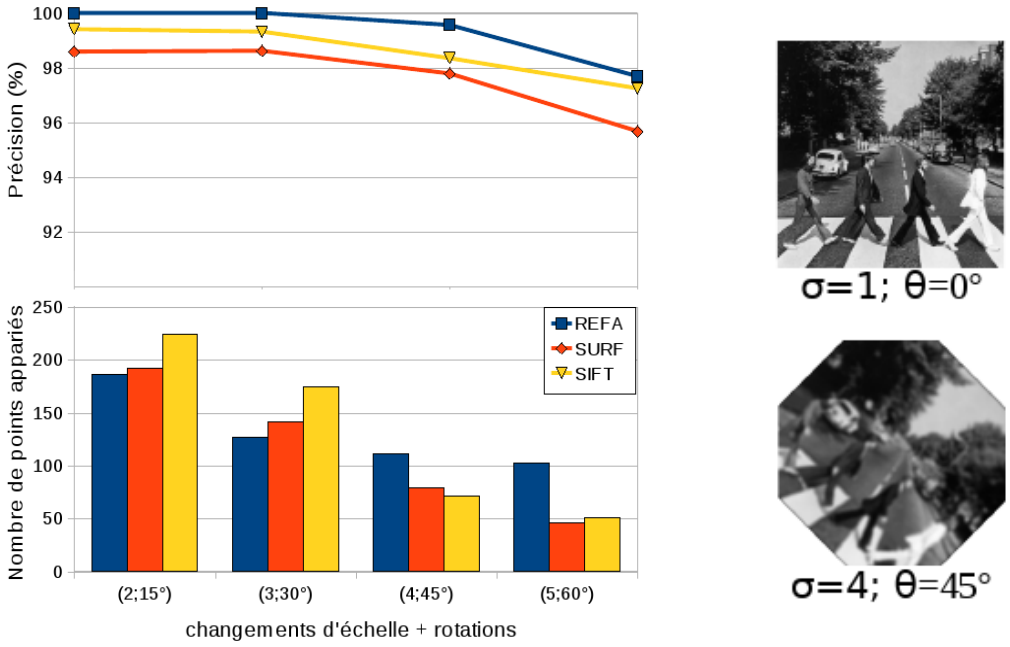


FIG. F.10 – Précision et nombre de points appariés pour un couplages changements d'échelle/rotations (Beatles).

– Transformations de type **étirements + rotations**.

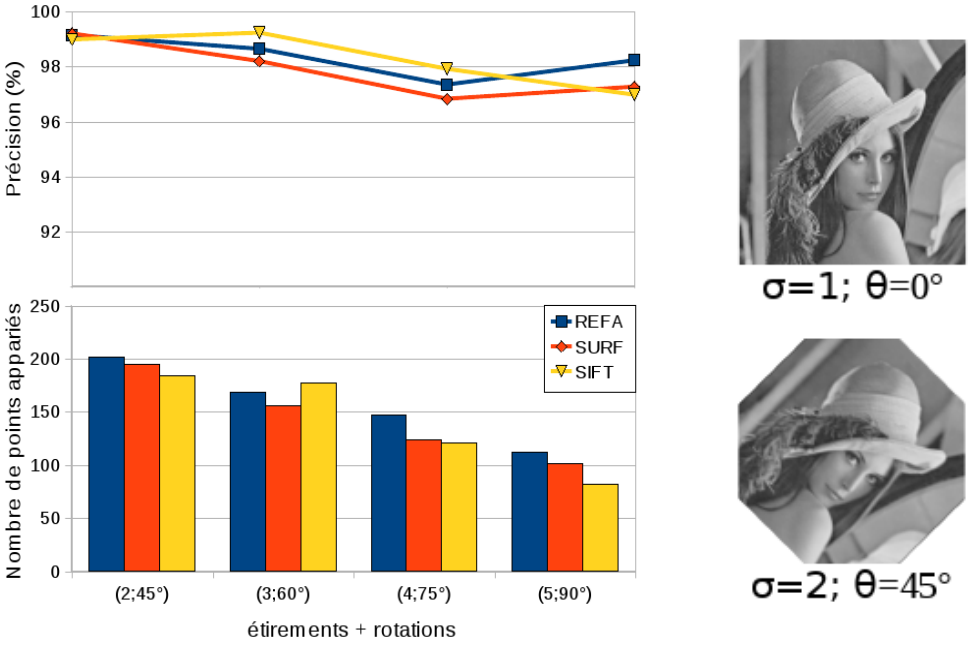


FIG. F.11 – Précision et nombre de points appariés pour un couplage étirements et rotations (Lena).

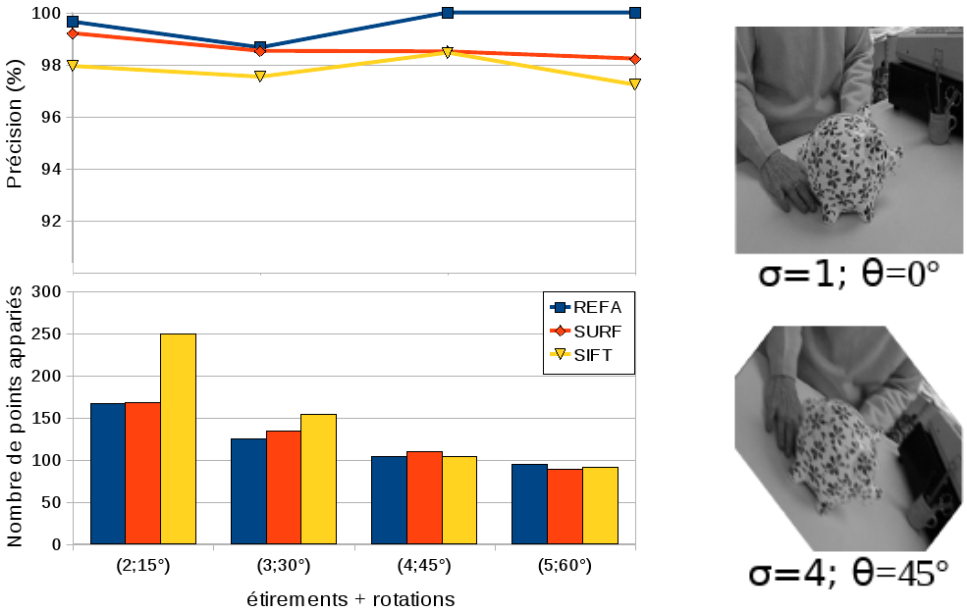


FIG. F.12 – Précision et nombre de points appariés pour un couplages étirements et rotations (Pig).

F.1.2 Transformations réelle

– Concernant les transformations composées d'un couplage **changements d'échelles et rotations**, les courbes de la figure F.13 présentent les résultats obtenus pour les images Boat.

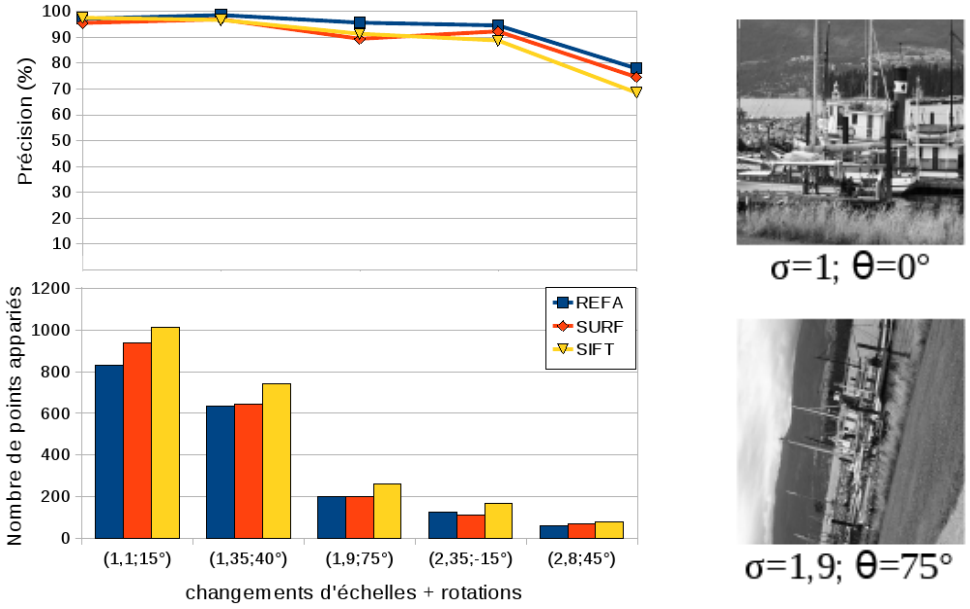


FIG. F.13 – Précision et nombre de points appariés pour un couplage changements d'échelle et rotations (Boat).

– Les courbes de la figure F.14 illustrent les résultats obtenus pour des de **modifications de compression JPEG** de l'image.

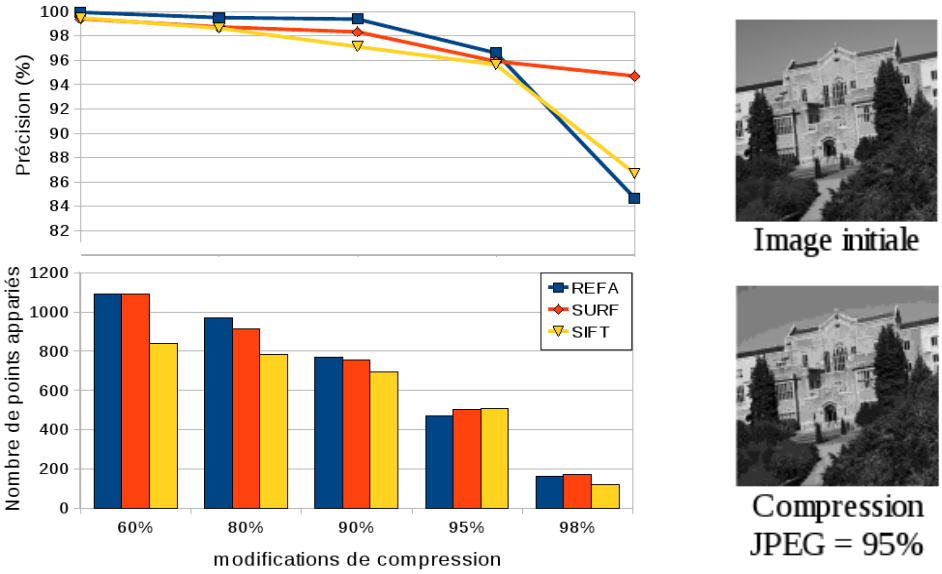


FIG. F.14 – Précision et nombre de points appariés pour des modifications de compression de l'image (images Ubc).

– Nous étudions des transformations de type **changements de luminosité**. Les courbes de la figure F.15 présentent les résultats obtenus pour les images Leuven.

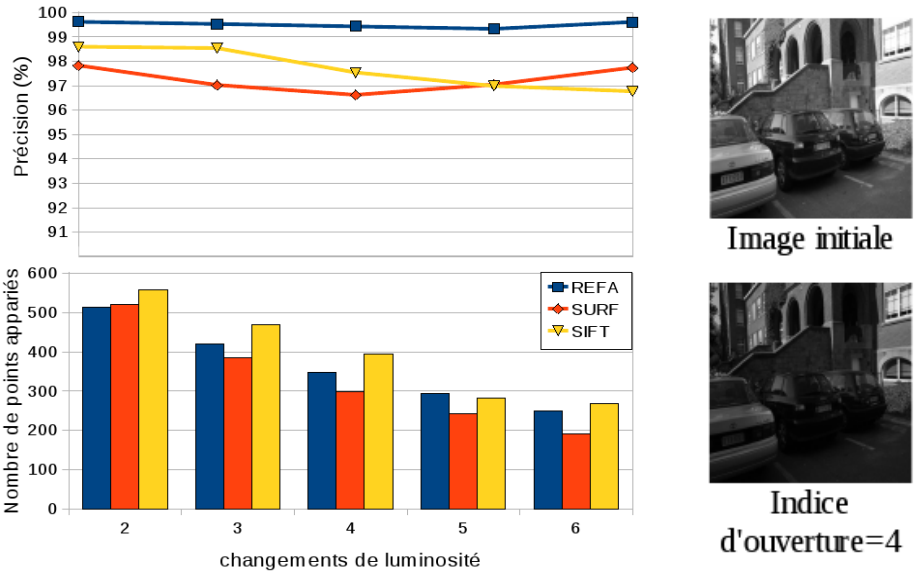


FIG. F.15 – Précision et nombre de points appariés en fonction de l'indice d'ouverture de la caméra (plus l'indice est élevé, plus l'ouverture est petite).

– Le **bruitage** de l'image peut entraîner de fortes détériorations des résultats, nous proposons donc d'en observer les conséquences (figure F.16).

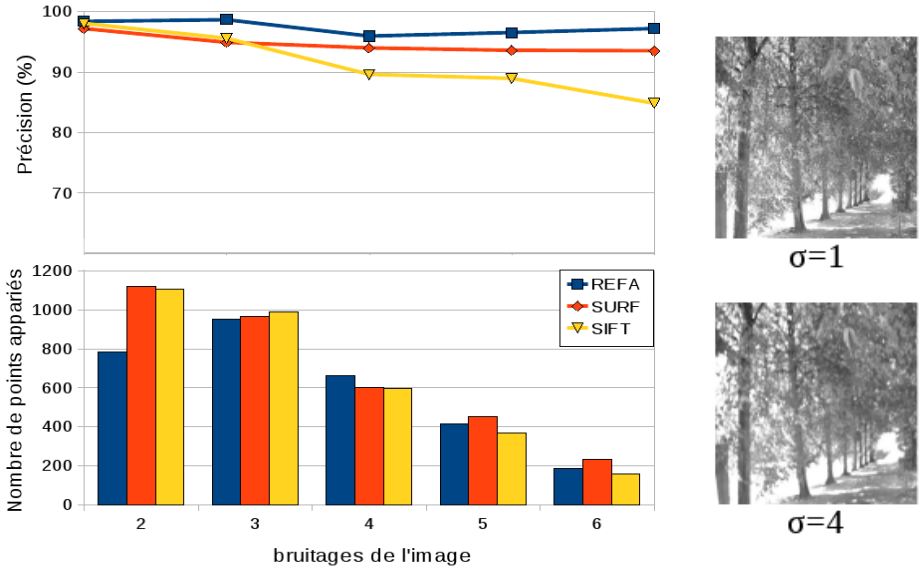


FIG. F.16 – Précision et nombre de points appariés pour des ajouts de bruits gaussiens dans l'image (écart type $\sigma \in [\![2;6]\!]$, images Trees).

– Nous proposons d'analyser un dernier type de transformations : les **changements de point de vue**. Les figures F.17 et F.18, présentent les résultats obtenus pour les images Wall et Graffiti.

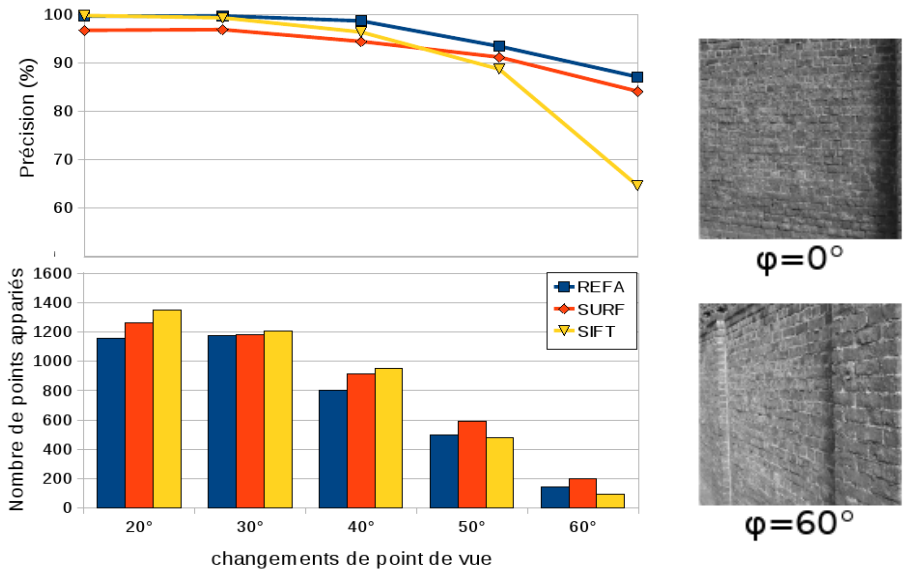


FIG. F.17 – Précision et nombre de points appariés pour des changements de point de vue (images Wall).

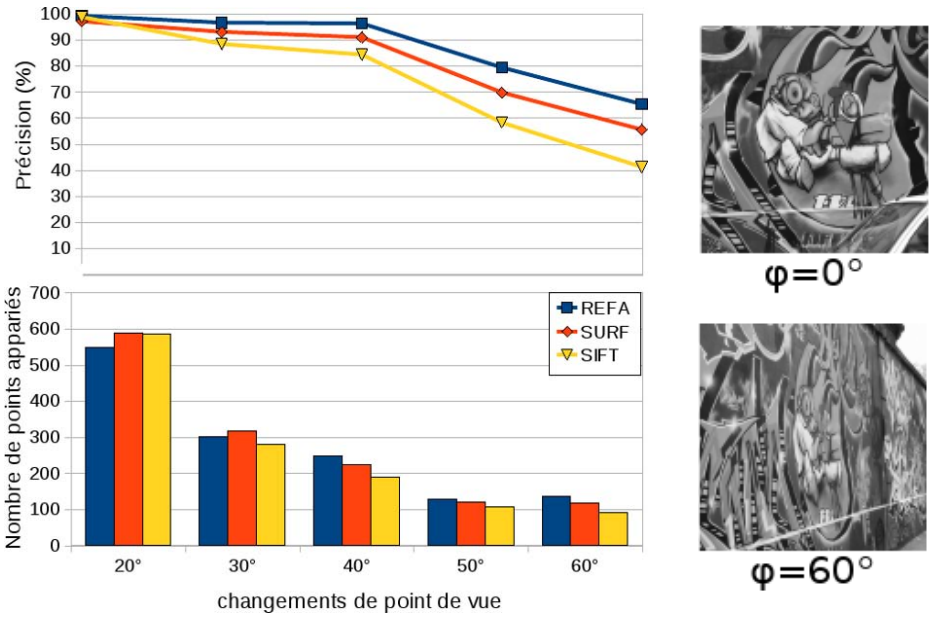


FIG. F.18 – Précision et nombre de points appariés pour des changements de point de vue (images Graffiti).

F.2 Ajout de tests pour la validation 2D+t de notre méthode

Pour le domaine spatio-temporel, nous avons également testé l'ensemble des transformations présentées dans notre manuscrit, en ajoutant un critère permettant de fournir un nombre initial de points détectés identique pour les trois méthodes. Il apparaît que les analyses et bilans énoncés dans notre manuscrit se confirment. En effet, les taux de précision du SIFT et du SURF augmentent, mais dans la majorité des cas, notre méthode conserve la meilleure précision. Le nombre de points appariés reste en moyenne assez proche de celui des deux autres méthodes et décroît, pour certaines transformations, plus lentement, caractérisant encore une fois une meilleure stabilité.

F.2.1 Transformations synthétiques

– Nous souhaitons analyser des transformations de type **translations**. Les figures F.19 et F.20 illustrent la précision obtenue et le nombre de points mis en correspondance pour les séquences 1 et 2.

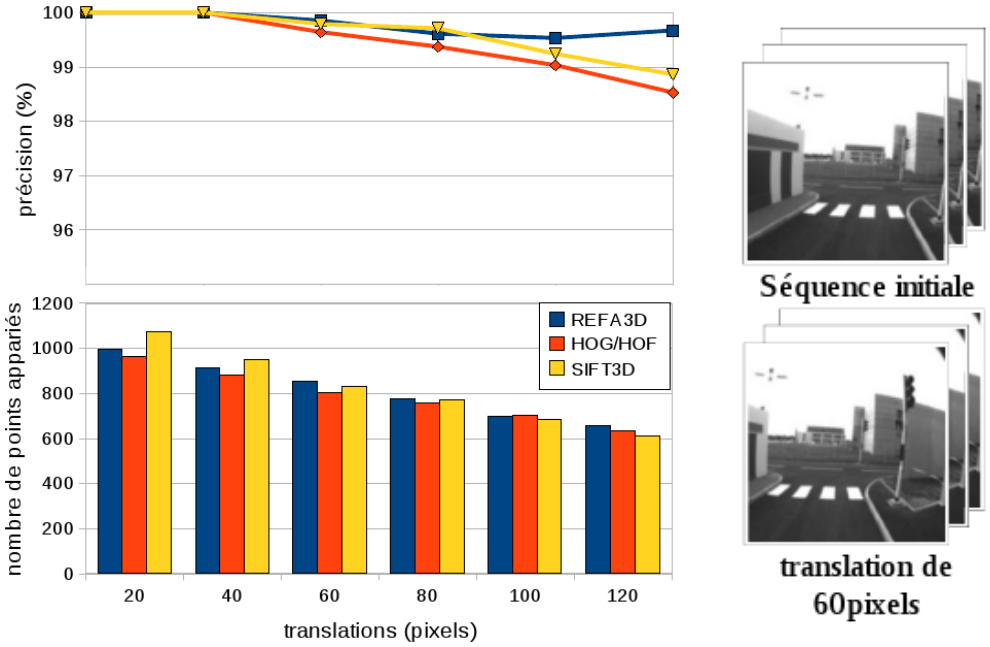


FIG. F.19 – Précision et nombre de points appariés pour des translations horizontales (séquence 1).

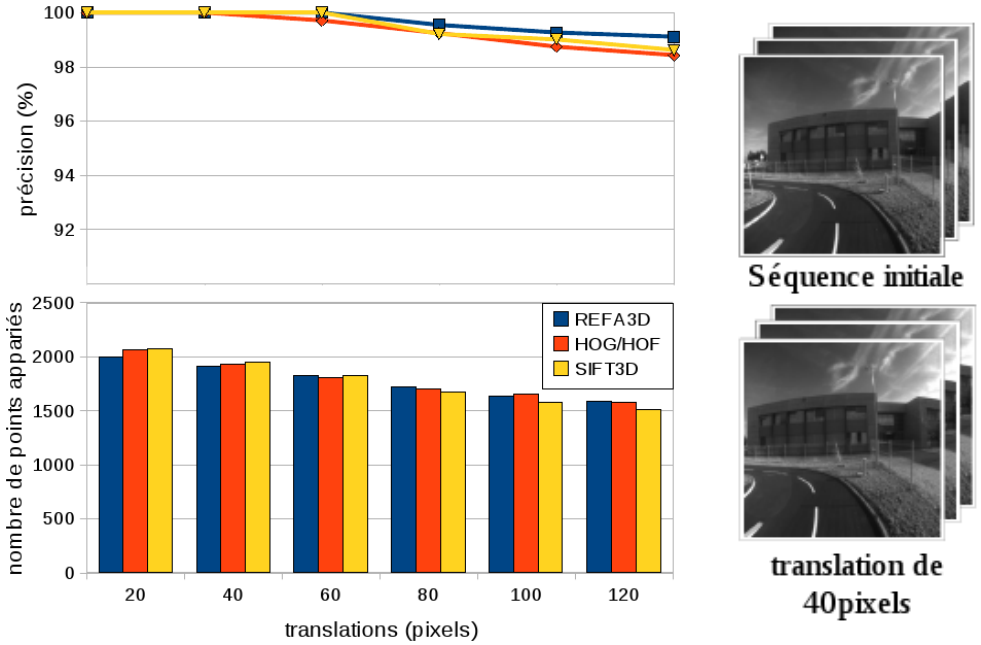


FIG. F.20 – Précision et nombre de points appariés pour des translations horizontales (séquence 2).

– Nous proposons également d'étudier l'influence de modifications de type **rotations** sur les performances de notre méthode. La figure F.21 regroupe les résultats obtenus, précision et nombre de points appariés, pour la séquence 2.

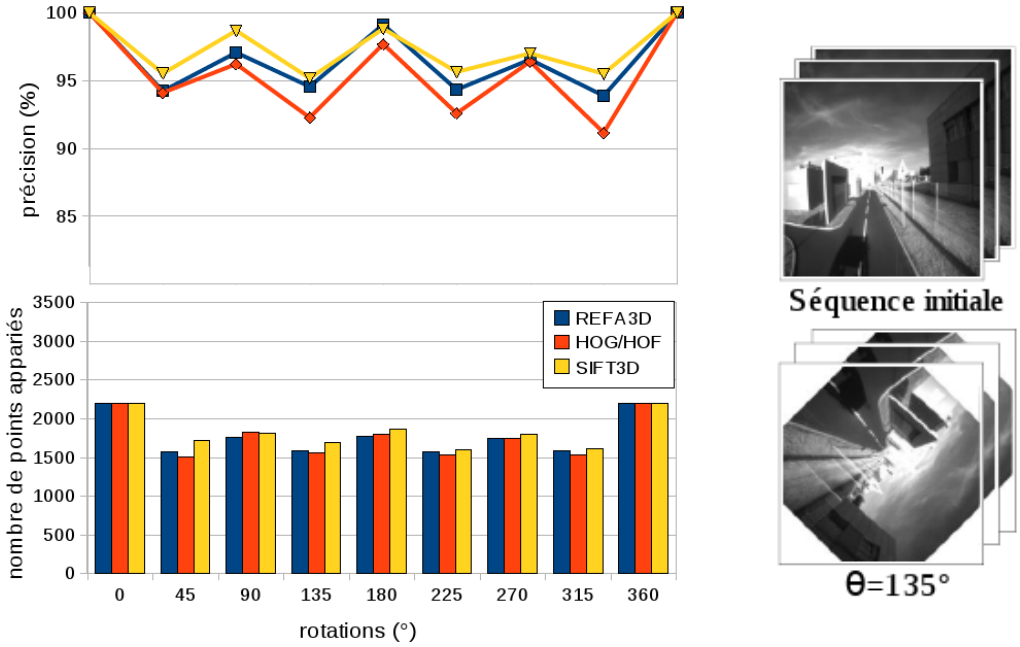


FIG. F.21 – Précision et nombre de points appariés pour des rotations (séquence 2).

– L'étude de **changements d'échelle spatiale** permet de mettre en avant l'utilité de l'adaptation des ellipsoïdes et de la qualité du paramètre σ extrait. Les figures F.22 et F.23 regroupent les résultats obtenus pour ce type de transformations.

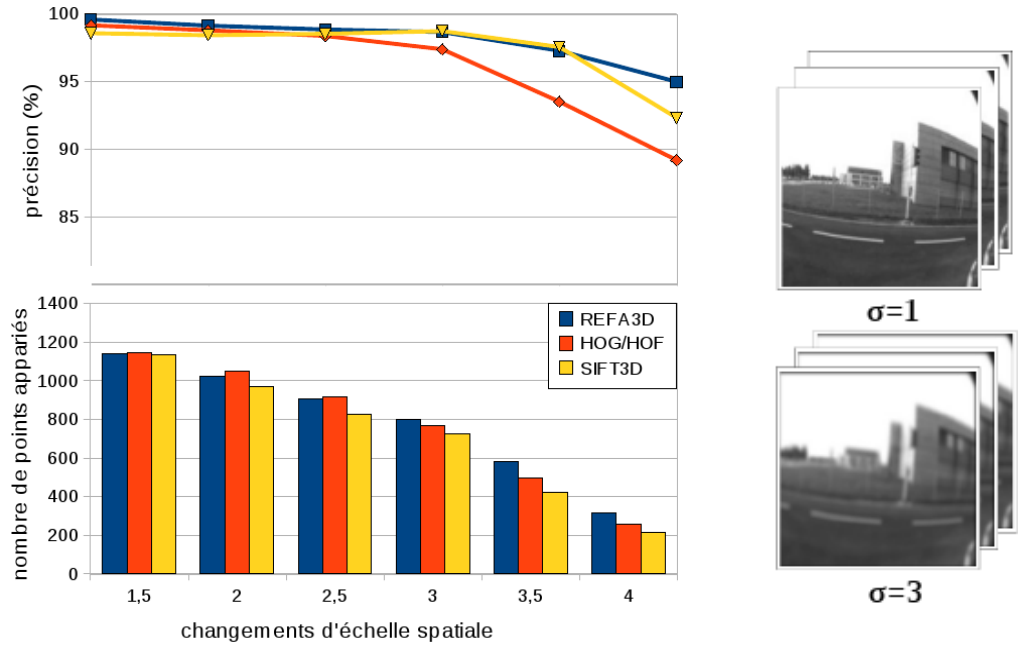


FIG. F.22 – Résultats pour des changements d'échelle spatiale (synthétique ; séquence 1).

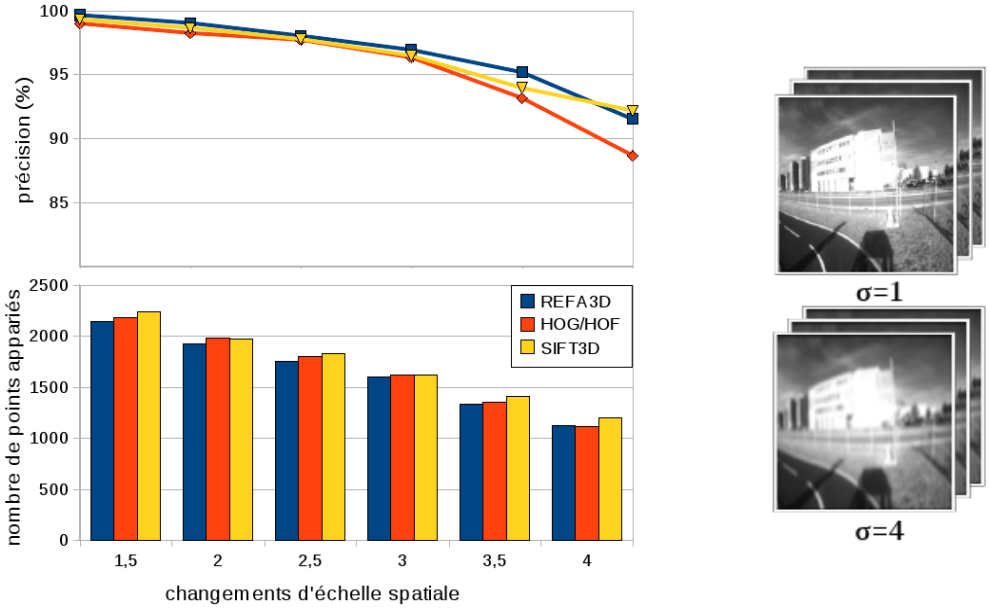


FIG. F.23 – Résultats pour des changements d'échelle spatiale (synthétique ; séquence 2).

– Nous proposons une dernière étude simulant une accélération du déplacement de la caméra. Les figures F.24 et F.25 illustrent les résultats pour ce type de transformations. Les absisces correspondent aux intervalles de sélection des images prélevées dans la séquences initiales (2 : 1 image sur 2, 3 : 1 image sur 3, ...).

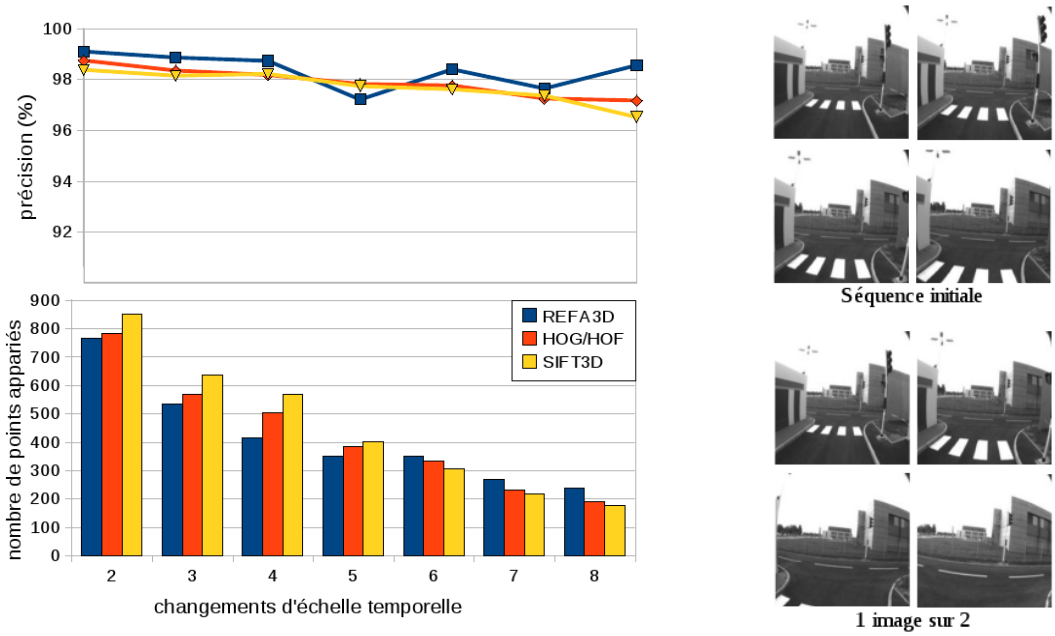


FIG. F.24 – Résultats pour des changements d'échelle temporelle (synthétique ; séquence 1).

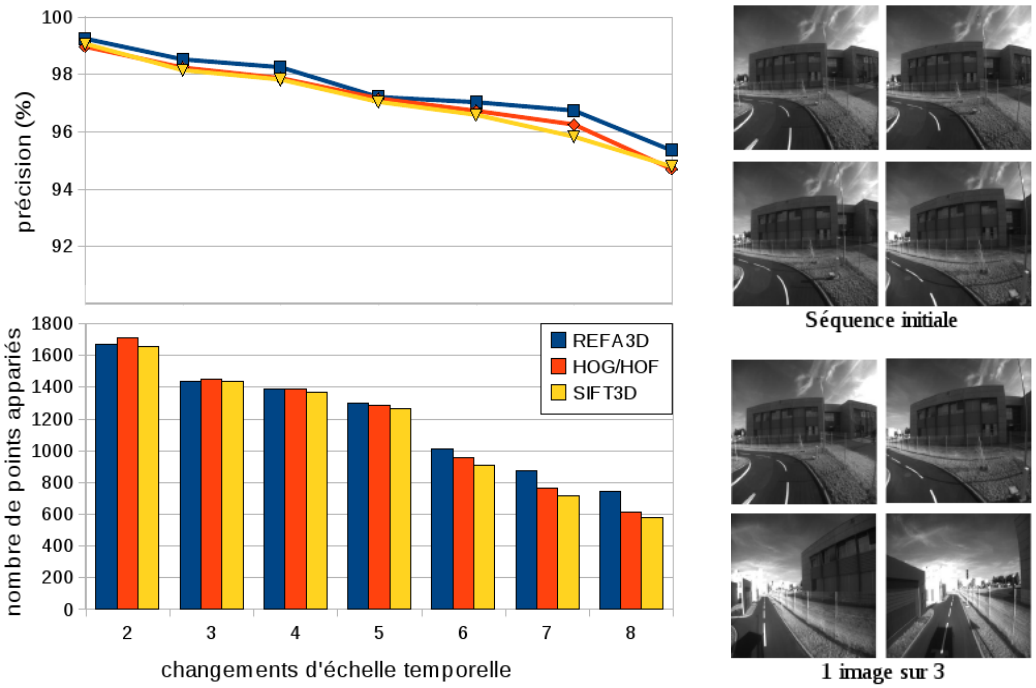


FIG. F.25 – Résultats pour des changements d'échelle temporelle (synthétique ; séquence 2).

F.2.2 Transformations réelle

Nous comparons notre approche aux SIFT3D et HOG/HOF suivant deux jeux de séquences. Le premier correspond à un quart de tour du rond-point de la plateforme PAVIN et le second s'appuie quant à lui sur le tour complet. Chaque jeu se compose de trois trajectoires (intérieure, centrée et extérieure) étudiées les unes par rapport aux autres et mises en correspondance. Ces différents tests permettent de mettre en avant la robustesse des méthodes étudiées vis à vis d'un ensemble de transformations pouvant s'apparenter à une dérive du véhicule. Les figures F.26 et F.27 illustrent les résultats obtenus.

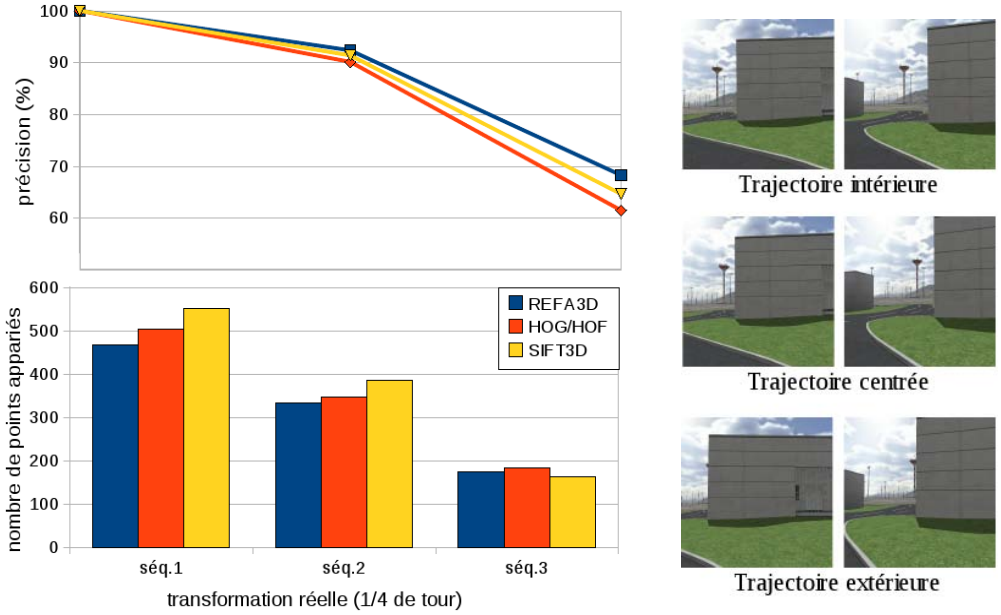


FIG. F.26 – Résultats des mises en correspondance des 3 trajectoires issues d'un quart de tour du rond-point (ASROCAM).

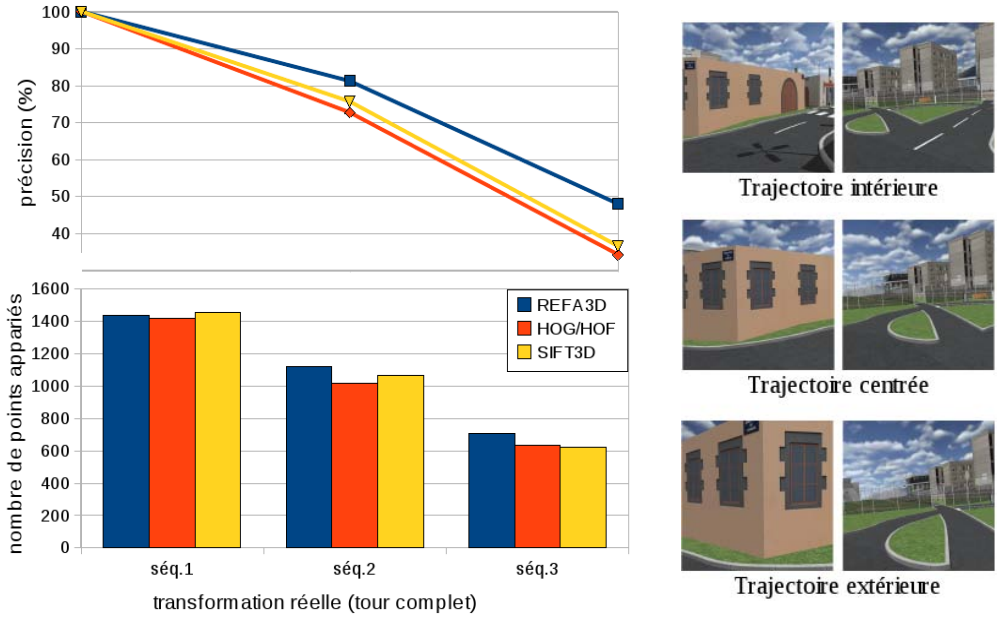


FIG. F.27 – Résultats des mises en correspondance des 3 trajectoires issues du tour complet du rond-point (ASROCAM).

Publications dans le cadre de cette thèse

Publication dans un chapitre de livre :

– REFA : a Robust E-HOG for Feature Analysis for local description of interest points, Manuel Grand-brochier, Christophe Tilmant, Michel Dhome • Communications in Computer and Information Science (CCIS 2011) publié par Springer-Verlag (sous presse).

Conférence internationale avec actes et comité de lecture :

– REFA3D : Robust Spatio-temporal analysis of video sequences, Manuel Grand-brochier, Christophe Tilmant, Michel Dhome • International Conference on Computer Vision Theory and Applications (VISAPP 2012) • Rome, Italie (en cours de soumission).

– Method of extracting interest points based on multi-scale detector and local E-HOG descriptor, Manuel Grand-brochier, Christophe Tilmant, Michel Dhome • International Conference on Computer Vision Theory and Applications (VISAPP 2011) • Faro, Portugal.

– Method of interest points characterization based on C-HOG local descriptor, Manuel Grand-brochier, Christophe Tilmant, Michel Dhome • International Symposium on Visual Computing (ISVC 2010) • Las-Vegas, Etats-Unis.

Conférence nationale avec actes et comité de lecture :

– Combinaison du détecteur de points fast-hessien avec un descripteur local basé C-HOG, Manuel Grand-brochier, Christophe Tilmant, Michel Dhome • Manifestation des Jeunes Chercheurs en Sciences et Technologies de l'information et de la Communication (Majecstic 2010) • Bordeaux, France.

– Descripteur local d'image invariant aux transformations affines, Manuel Grand-brochier, Christophe Tilmant, Michel Dhome • Congrès des jeunes chercheurs en vision par ordinateur (ORASIS 2009) • Trégastel, France.

Bibliographie

[1] C. Achard, E. Bigorgne, and J. Devars. A sub-pixel and multispectral corner detector. *International Conference on Pattern Recognition*, 3 :971–974, 2000.

[2] C. Achard, G. Mostafaoui, and M. Milgram. Object tracking with spatio-temporal blob. *Conference on Machine Vision Apllications*, pages 23–26, 2005.

[3] C. Achard, X. Qu, A. Mokhber, and M. Milgram. A novel approach for recognition of human actions with semi-global features. *Conference of Machine Vision and Apllications*, 19 :27–34, 2008.

[4] E. Adelson and J. Bergen. Spatiotemporal energy models for the perception of vision. *Journal of the Optical Society of America*, 2(2) :284–299, 1985.

[5] S. Arya and D. Mount. Approximate nearest neighbor queries in fixed dimensions. *ACM-SIAM*, pages 271–280, 1993.

[6] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu. An otimal algorithm for approximate nearest neighbor searching in fixed dimensions. *ACM-SIAM*, pages 573–582, 1994.

[7] H. Asada and M. Brady. The curvature primal sketch. *IEEE Pattern Analysis and Machine Intelligence*, 8(1) :2–14, 1986.

[8] P. Aschwanden and W. Guggenbühl. *Experimental Results from a Comparative Study on Correlation-Type Registration Algorithms*. Robust computer vision, 1992.

[9] J. Badri, C. Tilmant, J-M. Lavest, Q-C. Pham, and P. Sayd. Camera-to-camera mapping for hybrid pan-tilt-zoom sensors calibration. *Lecture Notes in Computer Science*, 4522 :132–141, 2007.

[10] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journalof Computer Vision*, 12(1) :43–77, 1994.

[11] J. Bauer, N. Sünderhauf, and P. Protzel. Comparing several implementations of two recently published feature detectors. *Intelligent Autonomous Vehicles*, 2007.

[12] A. Baumberg. Reliable feature matching across widely separated views. *IEEE Conference on Computer Vision and Pattern Recognition*, 1 :774–781, 2000.

[13] H. Bay, T. Tuylelaars, and L. Van Gool. Surf : Speeded up robust features. *European Conference on Computer Vision*, pages 404–417, 2006.

[14] A. Beaton and J. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16 :147–185, 1974.

[15] P. Beaudet. Rotationally invariant image operators. *International Journal of Current Pharmaceutical Research*, pages 579–586, 1978.

[16] S. Benhimane and E. Malis. Real-time image-based tracking of planes using efficient second-order minimization. *International Conference in Intelligent Robots Systems*, 1 :943–948, 2004.

[17] M. Bern, D. Eppstein, and S. Teng. Parallel construction of quadtrees and quality triangulations. *Workshop Algorithms Data Struct.*, 709 :188–199, 1993.

[18] M. Brown and D. Lowe. Unsupervised 3d object recognition and reconstruction in unordered datasets. *International Conference on 3-D Digital Imaging and Modeling*, pages 56–63, 2005.

[19] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief : Binary robust independent elementary features. *European Conference on Computer Vision*, 2010.

[20] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod. Chog : Compressed histogram of gradients. a low bit-rate feature descriptor. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2504–2511, 2009.

[21] Y. Chen and Y. Hung. Feature-based displacement field estimation for visual tracking by using coarse-to-fine block matching. *International Conference on Artificial Intelligence*, pages 129–136, 1996.

[22] H. Cheng, Z. Liu, N. Zheng, and J. Yang. A deformable local image descriptor. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[23] W. Cheung and G. Hamarneh. n-sift : N-dimensional scale invariant feature transform. *IEEE Transactions on Image Processing*, 18(9) :2012–2021, 2009.

[24] H. Chiu and T. Lozano-Perez. Matching interest points using affine invariant concentric circles. *International Conference on Pattern Recognition*, 2 :167–170, 2006.

[25] A. Choksuriwong, H. Laurent, and B. Emile. Etude comparative de descripteur invariants d'objets. *ORASIS*, 2005.

[26] A. Choksuriwong, H. Laurent, and B. Emile. Object recognition using local characterisation and zernike moments. *Advanced Concepts for Intelligent Vision Systems*, 3708 :108–115, 2005.

[27] A. Choksuriwong, H. Laurent, B. Emile, and C. Rosenberger. Comparative study of global invariant descriptors for object recognition. *Journal of Electronic imaging*, 17(2), 2008.

[28] C-W. Chong, P. Raveendran, and R. Mukundan. A comparative analysis of algorithms for fast computation of zernike moment. *Pattern Recognition*, 36 :731–742, 2003.

[29] K. Clarkson. Fast algorithms for the all nearest neighbors problem. *IEEE Symposium on the Foundations of Computer Science*, pages 226–232, 1983.

[30] K. Clarkson. An algorithm for approximate closest-point queries. *ACM Symposium Computer Geometry*, pages 160–164, 1994.

[31] J. Crowley and A. Parker. A representation for shape based on peaks and ridges in the difference of low pass transform. *IEEE Transformations on Pattern Analysis and Machine Intelligence*, 6(2) :156–170, 1984.

[32] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[33] P. Delmas. *Génération active des déplacements d'un véhicule agricole dans son environnement*. PhD thesis, Université Blaise Pascal - Clermont II, 2011.

[34] R. Deriche and G. Giraudon. A computational approach for corner and vertex detection. *International Journal of Computer Vision*, 10(2) :101–124, 1993.

[35] S. Derrode, M. Daoudi, and F. Ghorbel. Invariant content-based image retrieval using a compete set of fourier-mellin descriptors. *IEEE International Conference on Multimedia Computing and Systems*, 2 :877–881, 1999.

[36] S. Derrode and F. Ghorbel. Robust and efficient fourier-mellin transform approximations for gray-level image reconstruction and complete invariant description. *Computer Vision and Image Understanding*, 83(1) :57–78, 2001.

[37] R. Dinesh and G. Guru. Mathematical morphology based corner detection scheme : a non-parametric approach. *Electronics and Communication Sciences Unit. Indian Statistical Institute*, 2004.

[38] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *IEEE International Conference on Computer Vision*, 2005.

[39] L. Dreschler and H. Nagel. Volumetric model and 3d trajectory of a moving car derived from monocular tv frame sequences of a street scene. *Computer Graphics ans Image Processing*, 20 :199–228, 1982.

[40] J. Falcou and J. Serot. Application of template-based metaprogramming compilation techniques to the efficient implementation of image processing algorithms on simd-capable processors. *Advanced Concepts for Intelligent Vision Systems*, 2004.

[41] J. Falcou, J. Serot, T. Chateau, and J.T Lapreste. Real time parallel implementation of a particle filter based visual tracking. *Computation Intensive Methods for Computer Vision*, 2006.

[42] N. Fischler and R. Bolles. Random sample consensus : A paradigm for model fitting with application to image analysis and autimated cartography. *Communication of the ACM*, 24 :381–395, 1981.

[43] J. Friedman, J. Bentley, and R. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3 :209–226, 1977.

[44] F. Ghorbel. A complete invariant description for gray-level images by the harmonic analysis approach. *Pattern Recognition Letters*, 15 :1043–1051, 1994.

[45] V. Gouet, P. Montesinos, R. Deriche, and D. Pelé. Evaluation de détecteurs de points d'intérêt pour la couleur. *Reconnaissance des Formes et Intelligence Artificielle*, pages 257–266, 2000.

[46] M. Grabner, H. Grabner, and H. Bischof. Fast approximated sift. *Asian Conference on Computer Vision*, pages 918–927, 2006.

[47] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, pages 147–151, 1988.

[48] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. *Cambridge University Press*, 2, 2004.

[49] W. He, T. Yamashita, H. Lu, and S. Lao. Surf tracking. *International Conference on Computer Vision*, pages 1586–1592, 2009.

[50] Y. Hel-Or and H. Hel-Or. Real-time pattern matching using projection kernels. *IEEE Pattern Analysis and Machine Intelligence*, 27(9) :1430–1445, 2005.

[51] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17 :185–204, 1981.

[52] P. Hough. Machine analysis of bubble chamber pictures. *International conference on High Energy Accelerators and Instrumentation*, pages 554–556, 1959.

[53] M. Hu. Visual pattern recognition by moment invariants. *IEEE Transactions Information Theory*, 8 :179–187, 1962.

[54] P. Huber. Robust statistics. *Wiler*, 1981 (seconde édition : 2004). ISBN : 0470129905.

[55] R. Hummel and S. Zucker. On the fundations of relaxation labeling processes. *IEEE Pattern Analysis and Machine Intelligence*, 5(3) :267–287, 1983.

[56] J. Illingworth and J. Kittler. A survey of the hough transform. *Computer Vision, Graphics, and Image Processing*, 44 :87–116, 1988.

[57] L. Juan and O. Gwun. A comparison of sift, pca-sift and surf. *International Journal of Image Processing*, 3(4) :143–152, 2009.

[58] F. Jurie and C. Schmid. Scale-invariant shape features for recognition of onject categories. *IEEE Computer Vision and Pattern Recognition*, 2 :90–96, 2004.

[59] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. *International Conference on Computer Vision*, 1 :166–173, 2005.

[60] L. Kitchen and A. Rosenfeld. Gray level corner detection. *Pattern Recognition Letters*, 1 :95–102, 1982.

[61] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. *British Machine Vision Conference*, pages 995–1004, 2008.

[62] J. Koenderink. The structure of images. *Biological Cybernetics*, pages 363–370, 1984.

[63] I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg. Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, 108(3) :207–229, 2007.

[64] I. Laptev and T. Lindeberg. Space-time interest points. *IEEE International Conference on Computer Vision*, 1 :432–439, 2003.

[65] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. *Computer and Information Science*, 3667 :91–103, 2006.

[66] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77 :259–289, 2005.

[67] T. Lindeberg. Scale-space for discret signals. *IEEE Transations on Pattern Analysis and Machine Intelligence*, pages 234–254, 1990.

[68] T. Lindeberg. Discrete derivative approximations with scale-space properties : A basis for low-level feature extraction. *Journal of Mathematical Imaging and Vision*, pages 349–376, 1993.

[69] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994. ISBN 0-7923-9418-6.

[70] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2) :79–116, 1998.

[71] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-d depth cues from affine distortions of local 2-d structure. *Image and Vision Computing*, 15(6) :415–434, 1997.

[72] H. Liu and J. Zhou. Motion planning for human-robot interaction based on stereo vision and sift. *International Conference on Systems, Man and Cybernetics*, pages 830–834, 2009.

[73] D. Lowe. Object recognition from local scale-invariant features. *IEEE International Conference on Computer Vision*, pages 1150–1157, 1999.

[74] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004.

[75] F. Malartre. *Perception intelligente pour la navigation rapide de robots mobiles en environnement naturel*. PhD thesis, Université Blaise Pascal - Clermont II, 2011.

[76] E. Malis and E. Marchand. Méthodes robustes d'estimation pour la vision robotique. *Journées Nationales de la Recherche en Robotique*, 1 :51–59, 2005.

[77] L. Masson, M. Dhome, and F. Jurie. Tracking 3d objects using flexible models. *British Machine Vision Conference*, pages 1–10, 2005.

[78] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. *British Machine Vision Conference*, pages 384–396, 2002.

[79] J. Mennesson, C. Saint-Jean, and L. Mascarilla. De nouveaux descripteurs de fourier géométriquqes pour l'analyse d'images couleur. *RFIA*, pages 599–606, 2010.

[80] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. *IEEE International Conference on Computer Vision*, pages 525–531, 2001.

[81] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. *European Conference on Computer Vision*, 1 :128–142, 2002.

[82] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 1(60) :63–86, 2004.

[83] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Pattern Analysis and Machine Intelligence*, 27(10) :1615–1630, 2005.

[84] K. Mikolajczyk, Tuytelaars, C. Schmid, Zisserman, Matas, Schaffalitzky, Kadir, and Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65 :43–72, 2006.

[85] H. Moravec. Towards automatic visual obstacle avoidance. *International Joint Conferences on Artificial Intelligence*, page 584, 1977.

[86] J-M. Morel and G. Yu. Asift : A new framework for fully affine invariant image comparaison. *SIAM Journal on Imaging Sciences*, 2(2) :438–469, 2009.

[87] J-M. Morel and G. Yu. A fully affine invariant image comparison method. *International Conference on Acoustics, Speech and Signal Processing*, pages 1597–1600, 2009.

[88] H. Nagel. Displacement vectors derived from second order intensity variations in image sequences. *Computer Vision, Graphics and Image Processing*, 21 :85–117, 1983.

[89] J. Noble. Finding corners. *Image and Vision Computing*, 6 :121–128, 1988.

[90] K. Peng, X. Chen, D. Zhou, and Y. Liu. 3d reconstruction based on sift and harris feature points. *International Conference on Robotics and Biomimetics*, pages 960–964, 2010.

[91] F. Porikli. Integral histogram : A fast way to extract histograms in cartesian spaces. *IEEE Conference on Computer Vision and Pattern Recognition*, 1 :829–836, 2005.

[92] J. Rabin, J. Delon, Y. Gousseau, and L. Moisan. Mac-ransac : a robust algorithm for the recognition of multiple objects. *International Symposium on 3D Data Processing, Visualization and Transmission*, 2010.

[93] Y. Raoui, E.H. Bouyakhf, M. Devy, and F. Regragui. *Global and Local Descriptors for Content Based Image Retrevial anad Object Recognition*, volume 5. Applied Mathematical Sciences, 2011. ISSN 1312-885X.

[94] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. *European Conference on Computer Vision*, pages 430–443, 2006.

[95] P. Rousseeuw. Least median of squares regression. *Journal American Statistic Association*, 79 :871–880, 1984.

[96] P. Rousseeuw and A. Leroy. Robust regression and outlier detection. *John Wiley and Sons*, 1987.

[97] R. Rusu, J. Bandouch, F. Meier, I. Essa, and M. Beetz. Human action recognition using global point feature histograms and action shapes. *Advanced Robotics journal, Robotics Society of Japan*, 2009.

[98] B. Schiele and A. Waibel. Gaze tracking based on face-color. *International Workshop on Automatic Face and Gesture Recognition*, pages 344–349, 1995.

[99] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. *ACM Multimedia*, 2007.

[100] N. Shao, H. Li, L. Liu, and Z. Zhang. Stereo vision robot obstacle detection based on the sift. *WRI Global Congress on Intelligent Systems*, 2 :274–277, 2010.

[101] D. Shen, J. Lee, S. Kil, J. Ryu, E. Lee, and S. Hong. 3d reconstruction of scale-invariant features for mobile robot localization. *International Journal of Computer Science and Network Security*, 6 :101–109, 2006.

[102] J. Shi and C. Tomasi. Good features to track. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[103] D. Sidibe, P. Montesinos, and S. Janaqi. Mise en correspondance robuste d'invariants locaux par relaxation. *ORASIS*, 2007.

[104] M. Smith and J. Brady. Susan - a new approach to low level image processing. *International Journal of Computer Vision*, 23 :45–78, 1997.

[105] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism : Exploring photo collections in 3d. *ACM Transactions on Graphics*, pages 835–846, 2006.

[106] M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7 :11–32, 1991.

[107] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[108] Y. Tsai, Q. Wang, and S. You. Cdikp : A highly-compact local feature descriptor. *International Association for Pattern Recognition*, pages 1–4, 2008.

[109] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *IEEE Conference on Computer Vision and Pattern Recognition*, 1 :511–518, 2001.

[110] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. *British Machine Vision Conference*, 2009.

[111] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. *European Conference on Computer Vision*, 5303(2) :650–663, 2008.

[112] S. Di Zenzo. A note on the gradient of a multi-image. *Computer Vision Graphics and Image Processing*, 33 :116–125, 1986.

[113] X. Zhang and D. Zhao. A morphological algorithm for detecting dominant points on digital curves. *SPIE Proc. Non Linear Image Processing*, 2424 :372–383, 1995.

[114] G. Zhao, L. Chen, G. Chen, and J. Yuan. Kpb-sift : a compact local feature descriptor. *Association of Computing Machinery Multimedia*, pages 1175–1178, 2010.

[115] H. Zhou, Y. Yuan, and C. Shi. Object tracking using sift features and mean shift. *Computer Vision and Image Understanding*, 113 :345–352, 2009.

[116] O. Zuniga and R. Haralick. Corner detection using the facet model. *IEEE Conference on Computer Vision Pattern Recognition*, pages 30–37, 1983.