Mise en correspondance d'images à différentes résolutions à l'aide d'invariants aux paramètres intrinsèques

Matching images at different resolutions using intrinsics-free measures

S. Benhimane E. Malis

INRIA, Sophia Antipolis, Projet ICARE

2004, route des Lucioles - B.P. 93,
06902 Sophia Antipolis Cedex, France.
{Selim.Benhimane,Ezio.Malis}@sophia.inria.fr

Résumé

Cet article s'intéresse à la mise en correspondance entre points caractéristiques extraits à partir de deux images prises à des résolutions différentes par une caméra munie d'un objectif à focale variable. La caméra est supposée être stationnaire ou n'ayant effectué qu'un faible déplacement entre les deux images. Étant donné que les méthodes basées sur des données photométriques sont très peu fiables dans le cas d'une forte variation de la focale, nous proposons une nouvelle méthode d'appariement basée sur des mesures invariantes aux changements de paramètres intrinsèques. La mise en correspondance peut être utilisée dans plusieurs applications telles que la vidéo-surveillance, la vision active ou l'asservissement visuel.

Mots Clef

Vision 3D et géométrie, calibration, mise en correspondance, reconstruction, vision dynamique, vision active.

Abstract

This paper deals with matching between points of interest detected from two images at different resolutions taken using a camera equipped with a motorized zoom. The camera is supposed to be stationary or to be slightly displaced between the two images. In this context, the methods based on photometric data are inefficient, that is why we present a matching algorithm based on measures that are independent on the intrinsic parameters of the camera. The matching can be used in different applications such as video-surveillance, active vision or visual servoing.

Keywords

3D vision et geometry, calibration, matching, reconstruction, dynamic vision, active vision.

1 Introduction

La mise en correspondance est très utilisée dans le domaine de la vision par ordinateur. En effet, c'est l'une des étapes les plus importantes lors de la reconstruction 3D à partir d'une paire stéréo ou d'une séquence d'images ou lors de la détermination de la structure à partir d'un mouvement. C'est pour cette raison que beaucoup de chercheurs se sont intéressés à ce sujet durant les deux dernières décennies. Mais, malgré tout, la mise en correspondance demeure un exercice difficile et un sujet de recherche toujours ouvert. Les méthodes d'appariement classiques sont basées sur des grandeurs géométriques et/ou sur des grandeurs photométriques. Une approche de mise en correspondance [16] consiste à extraire les points d'intérêt à l'aide d'un filtre détecteur du type [5] ou [3], apparier ces points en se basant sur la mesure de la corrélation des fenêtres centrées en ces points, calculer la géométrie épipolaire à l'aide d'une estimation robuste de la matrice fondamentale du type [2] ensuite établir d'autres appariement grâce à de cette matrice. Ce genre de méthodes sont efficaces lors d'un faible mouvement entre les images et une faible variation de paramètres intrinsèques de la caméra. En effet, s'il y a une variation importante de paramètres intrinsèques et notamment de la focale, la corrélation du voisinage ne permet plus de donner un score de ressemblance fiable entre deux points caractéristiques. C'est pour cette raison que plusieurs études se sont intéressées à améliorer la mesure de la corrélation pour tenir compte

d'une variation plus importante dans l'image soit en estimant d'abord la transformation entre les deux images [10] soit en utilisant d'autres mesures de ressemblance entre les points tels que les invariants locaux [11]. Toutefois, ces méthodes demeurent inefficaces dans le cas d'une grande variation d'échelle entre les images. Certains auteurs proposent donc d'utiliser l'approche multi-échelle introduite par [15] et longuement étudiée dans [7] pour effectuer la mise en correspondance dans une pyramide d'images [4, 1]. Ces méthodes nécessitent le calcul des images à apparier sur plusieurs échelles et, dans le cas de [1], de tester la possibilité d'appariement de chacune des images initiales avec toutes les images de la pyramide formée avec l'autre image. Ceci coûterait très cher en temps de calcul et représenterait un inconvénient majeur dans certaines applications. Dans cet article, nous nous limitons à la mise en correspondance de deux images à différentes résolutions prises avec une caméra stationnaire ou n'ayant effectué qu'un faible déplacement entre les images. Nous proposons une approche différente basée principalement sur des mesures faites dans l'image. Ces mesures ont été introduites par [8] et sont invariantes aux paramètres intrinsèques de la caméra. Nous utilisons, également, des données photométriques pour effectuer notre appariement en tenant compte du fait que celles-ci sont très sensibles à la variation d'échelle. Contrairement à certains algorithmes classiques de mise en correspondance, nous ne cherchons pas à estimer la variation des paramètres (intrinsèques et extrinsèques) et les correspondances des points un à un, mais un appariement entre ensembles de points répétés dans les deux images. Si l'on arrive à savoir approximativement quels points ont été répétés dans les deux images alors l'appariement des points un à un est plus facile. Notre méthode diffère des autres approches de mise en correspondance par le fait que, à chaque itération, nous éliminons un ensemble de points qui a été extrait dans l'une des deux images et pas dans l'autre.

2 Modélisations

Nous considérons dans cet article l'espace comme étant cartésien Nous supposons que le repère absolu coïncide avec le repère de la caméra. Un point de l'espace $\boldsymbol{\mathcal{X}}_j \in \mathbb{P}^3$ est projeté sur un plan virtuel parallèle au plan $(\vec{x},\vec{y})$ en un point $\mathbf{m}_j = (x_j,y_j,1) \in \mathbb{P}^2$ tel que :

$$\mathbf{m}_j \propto \left[\begin{array}{cc} \mathbf{I}_{3\times 3} & \mathbf{0}_{3\times 1} \end{array} \right] \boldsymbol{\mathcal{X}}_j \qquad (1)$$

Cependant, le point obtenu en utilisant une caméra qui vérifie le modèle sténopé, donc qui réalise une projection perspective des points 3D, n'est pas $\mathbf{m}_j$ mais un point image $\mathbf{p}_{ij} = (u_{ij},v_{ij},1)$:

$$\mathbf{p}_{ij} = \mathbf{K}_i \mathbf{m}_j \qquad (2)$$

où $\mathbf{K}_i$ est la matrice des paramètres intrinsèques de la caméra :

$$\mathbf{K}_i = \left[\begin{array}{ccc} f_i & f_i s_i & u_{0i} \\ 0 & f_i r_i & v_{0i} \\ 0 & 0 & 1 \end{array} \right] \qquad (3)$$

où f_i est la focale exprimée en pixels, s_i modélise l'effet de cisaillement dû au défaut d'orthogonalité des axes du repère image et r_i est le rapport des dimensions d'un pixel, (u_{0i},v_{0i}) sont les coordonnées du point principal. Les paramètres intrinsèques de la caméra varient donc lors d'un zoom. Il faut remarquer que lors d'une variation du zoom, les paramètres extrinsèques varient également en fonction de la qualité de l'objectif [6]. Toutefois, pour chaque matrice $\mathbf{K}_i$ le modèle sténopé reste valide localement.

3 Les invariants aux paramètres intrinsèques

Les invariants utilisés dans cet article sont définis dans [8]. Ce sont des mesures dans l'espace image invariantes aux paramètres intrinsèques d'une caméra vérifiant le modèle sténopé. Supposons que nous avons n points ($n > 3$) extraits d'une scène. Ces points se projettent sur le plan virtuel en $\mathbf{m}_j$ suivant l'équation (1) et les points correspondants $\mathbf{p}_{ij}$ dans l'image sont obtenus à l'aide de l'équation (2). Soient les matrices (3×3) :

$$\mathbf{S}_{pi} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{p}_{ij} \mathbf{p}_{ij}^{\top} \quad \text{et} \quad \mathbf{S}_m = \frac{1}{n} \sum_{j=1}^{n} \mathbf{m}_j \mathbf{m}_j^{\top} \qquad (4)$$

La matrice $\mathbf{S}_{pi}$ est une matrice symétrique positive qui peut être calculée à partir des points images. La matrice $\mathbf{S}_m$ est également symétrique positive mais ne peut pas être calculée à partir de l'image. La décomposition de Cholesky de ces deux matrices donne :

$$\mathbf{S}_{pi} = \mathbf{T}_{pi} \mathbf{T}_{pi}^{\top} \quad \text{et} \quad \mathbf{S}_m = \mathbf{T}_m \mathbf{T}_m^{\top}$$

où $\mathbf{T}_m$ et $\mathbf{T}_{pi}$ sont des matrices triangulaires supérieures non singulières. D'après l'équation (2) et d'après l'unicité de la décomposition de Cholesky, on peut déduire la relation suivante entre ces deux matrices :

$$\mathbf{T}_{pi} = \mathbf{K}_i \mathbf{T}_m \qquad (5)$$

où la matrice $\mathbf{T}_m$ est indépendante des paramètres intrinsèques de la caméra $\mathbf{K}_i$. La matrice $\mathbf{T}_{pi}$ peut être utilisée pour définir la transformation projective suivante dans $\mathbb{P}^2$:

$$\mathbf{q}_{ij} = \mathbf{T}_{pi}^{-1} \mathbf{p}_{ij} \qquad (6)$$

où les points transformés $\mathbf{q}_{ij} \in \mathbb{P}^2$ sont invariants aux paramètres intrinsèques de la caméra. En effet, lors de

l'écriture de $\mathbf{q}_{ij}$, d'après (2) et (5), on a une simplification de la matrice $\mathbf{K}_i$:

$$\mathbf{q}_{ij} = \mathbf{T}_m^{-1}\mathbf{K}_i^{-1}\mathbf{K}_i\mathbf{m}_j = \mathbf{T}_m^{-1}\mathbf{m}_j$$

Par ailleurs, si l'on considère la distribution du nuage des points dans l'image en attribuant à tout les points la même masse, cette transformation a pour effet de normaliser certains moments d'ordre 2 de cette distribution et d'annuler les moments d'ordre 1 (voir annexe). Par exemple, considérons les deux images de la figure 1 prises par une caméra stationnaire munie d'un zoom motorisé. Les paramètres de la caméra correspondant aux deux images sont $\mathbf{K}_1$ et $\mathbf{K}_2$. Nous avons extrait manuellement 16 points dans chaque image. Les points $\mathbf{p}_{1j}$ de la première image (figure 1(a)) sont représentés par des croix jaunes et les points $\mathbf{p}_{2j}$ de la deuxième image (figure 1(b)) sont représentés par des cercles rouges. Il est important de souligner le fait que la mise en correspondance des points un à un n'est pas nécessaire pour calculer les invariants et qu'il suffit d'avoir le même ensemble de points extraits dans les deux images. Ayant ces deux ensembles, on calcule les matrices $\mathbf{T}_{p1}^{-1}$ et $\mathbf{T}_{p2}^{-1}$. Les invariants dans la première image sont :

$$\mathbf{q}_{1j} = \mathbf{T}_{p1}^{-1}\mathbf{p}_{1j} = \mathbf{T}_m^{-1}\mathbf{K}_1{}^{-1}\mathbf{K}_1\mathbf{m}_j = \mathbf{T}_m^{-1}\mathbf{m}_j \quad (7)$$

et ceux de la deuxième image sont :

$$\mathbf{q}_{2j} = \mathbf{T}_{p2}^{-1}\mathbf{p}_{2j} = \mathbf{T}_m^{-1}\mathbf{K}_2{}^{-1}\mathbf{K}_2\mathbf{m}_j = \mathbf{T}_m^{-1}\mathbf{m}_j$$

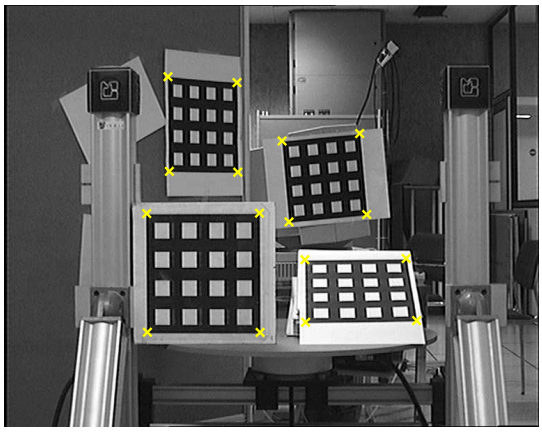
Il est évident que $\mathbf{A}\mathbf{q}_{1j} = \mathbf{A}\mathbf{q}_{2j}\ \forall j \in \{1,2,...,n\}$ et $\forall \mathbf{A}$ matrice (3×3) inversible. Si l'on choisit $\mathbf{A} = \mathbf{T}_{p1}$, il est possible de reprojeter les points d'une image dans l'échelle de l'autre.

$$\mathbf{p}_1 = \mathbf{T}_{p1}\mathbf{T}_{p2}^{-1}\mathbf{p}_2$$

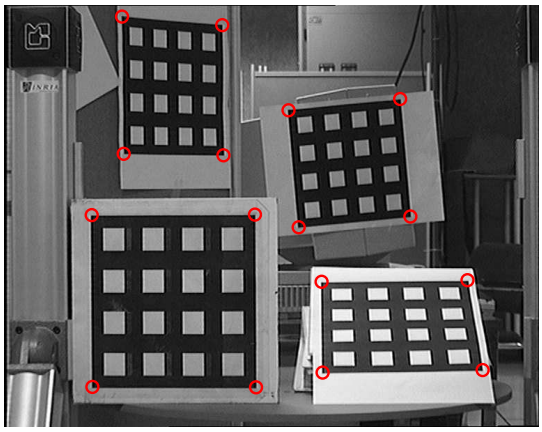
La figure 1(c) montre que, une fois la matrice $\mathbf{T}_{p1}\mathbf{T}_{p2}^{-1}$ a été bien estimée (i.e. les invariants $\mathbf{q}_{1j}$ et $\mathbf{q}_{2j}$ ont été correctement estimés), l'image 2 est parfaitement reprojetée dans l'échelle de l'image 1. Les points rouges extraits de l'image 2 ont été reprojetés en les points jaunes extraits dans l'image 1 (voir figure 1(b)).

4 Choix du détecteur des points caractéristiques

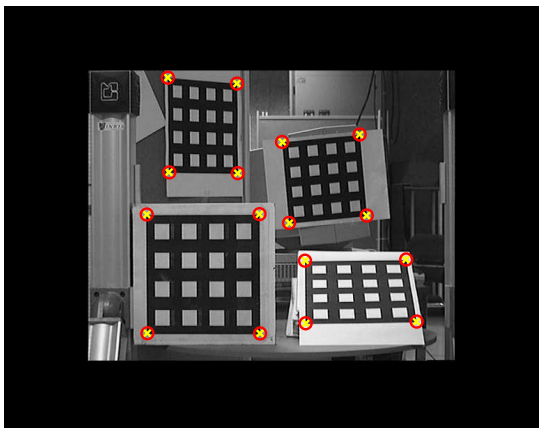
Pour effectuer une extraction automatique des points d'intérêt, nous utilisons le détecteur de Harris [5]. En effet, une étude comparative (effectuée dans [13]) entre plusieurs détecteurs permet d'affirmer que les propriétés de ce détecteur sont meilleures ou bien équivalentes aux autres détecteurs de point de vue répétabilité. Nous supposons que la répétabilité du détecteur vis-à-vis de la



(a) Image 1



(b) Image 2



(c) Reprojection de l'image 2 sur l'image 1

FIG. 1 – *Points extraits de deux images de résolutions différentes prises par une caméra effectuant un zoom.*

variation d'échelle est suffisante pour que le problème d'appariement reste bien posé. Nous rappelons les bases du détecteur de Harris. Tout d'abord, un point $\mathbf{p}_{ij}$ est considéré comme un point d'intérêt s'il est défini comme le centre d'une région pour laquelle une fonction d'intérêt est maximale en comparaison avec les régions voisines. Le détecteur de Harris utilise la fonction d'auto-corrélation comme fonction d'intérêt. On construit alors, une matrice $\mathbf{M}(\mathbf{p}_{ij})$:

$$\mathbf{M}(\mathbf{p}_{ij}) = \sum_{\mathbf{x}\in W} g(\mathbf{x}) \begin{bmatrix} I_u^2(\mathbf{x}) & I_u I_v(\mathbf{x}) \\ I_u I_v(\mathbf{x}) & I_v^2(\mathbf{x}) \end{bmatrix}$$

où W est une fenêtre centrée en $\mathbf{p}_{ij}$, $I_u = \frac{\partial I}{\partial u}(\mathbf{p}_{ij})$, $I_v = \frac{\partial I}{\partial v}(\mathbf{p}_{ij})$ et g est une gaussienne centrée en $\mathbf{p}_{ij}$. Si les valeurs propres de la matrice $\mathbf{M}(\mathbf{p}_{ij})$ sont proches $\lambda_1 \approx \lambda_2$ et sont supérieures à un certain seuil, alors $\mathbf{p}_{ij}$ est un point d'intérêt. Afin d'éviter le calcul des valeurs propres en chaque point, on utilise un score de détection. Dans la littérature, plusieurs scores ont été proposés. Dans [5], on propose :

$$score_1(\mathbf{p}_{ij}) = det - \alpha tr^2 \tag{8}$$

où det et tr sont le déterminant et la trace de la matrice $\mathbf{M}(\mathbf{p}_{ij})$ et α est un paramètre choisi dans $[0.04,0.06]$. Si le score du point $\mathbf{p}_{ij}$ est supérieur à un certain seuil (généralement pris à 10% du score le plus élevé sur tous les points de l'image) alors ce point est détecté comme un point d'intérêt. En effet, si $\lambda_1 \approx \lambda_2 \approx \lambda$ où λ est grand alors $score_1 \approx \lambda^2(1-4\alpha)$ et par conséquent le score est très grand. Si $\lambda_1 \gg \lambda_2$ alors $score_1 \approx \lambda_1(\lambda_2 - \alpha\lambda_1)$ et donc le score est faible. L'inconvénient de ce score est le choix du paramètre α qui nous paraît arbitraire et peu satisfaisant. Dans [3], on propose de calculer d'abord la trace de la matrice $\mathbf{M}(\mathbf{p}_{ij})$. Si la trace est inférieure à un certain seuil, alors le point $\mathbf{p}_{ij}$ appartient à une zone homogène. Sinon, il s'agit d'un point appartenant à une arête ou bien il s'agit d'un coin. Ensuite, on calcule le score suivant :

$$score_2(\mathbf{p}_{ij}) = \frac{4det}{tr^2} = \frac{4\lambda_1\lambda_2}{(\lambda_1+\lambda_2)^2} \tag{9}$$

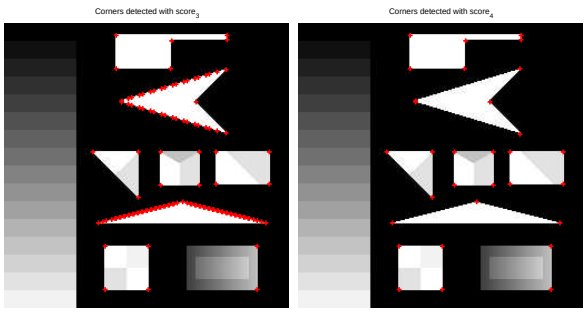
Si $\lambda_1 \approx \lambda_2$ alors $score_2 \approx 1$ et le point correspond à un coin. Si $\lambda_1 \gg \lambda_2$ alors $score_2 \approx 0$ et le point appartient à une arête. Cette méthode évite le choix arbitraire du paramètre α et permet d'avoir un score normalisé pour les coins. L'inconvénient de cette méthode est le choix de deux seuils de détection (l'un pour la trace et l'autre pour pour le score de détection). Dans [9], on propose de calculer le score :

$$score_3(\mathbf{p}_{ij}) = \frac{det}{tr} = \frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2} \tag{10}$$

Ce score permet de détecter en une seule étape les coins. En effet, si $\lambda_1 \approx \lambda_2 \approx \lambda$ où λ est grand (i.e. $\lambda > seuil$) alors $score_3 \approx \lambda_1/2$ et par conséquent le score est grand. Si $\lambda_1 \gg \lambda_2$ alors $score_1 \approx \lambda_2$ et donc le score est faible. L'inconvénient de cette méthode est qu'elle ne tient pas compte du cas où $\lambda_1 \gg \lambda_2 > seuil$. Ce cas peut arriver lorsque le point est sur une arête bien prononcée (voir figure 2(a)). Afin d'éviter ce phénomène, et pour améliorer la détection et avoir une extraction des coins uniquement, nous proposons un score qui n'est autre que le produit de (9) et de (10) :

$$score_4(\mathbf{p}_{ij}) = \frac{det}{tr} \times \frac{det}{tr^2} = \frac{(\lambda_1\lambda_2)^2}{(\lambda_1+\lambda_2)^3} \tag{11}$$

Le résultat de la détection est meilleur. En effet, en utilisant les mêmes paramètres de détection (même gaussienne et même méthode de calcul du seuil de détection), les points sur les arêtes et sur les côtés des triangles ne sont plus détectés (voir figure 2(b)).



(a) (b)

FIG. 2 – *Détection de points d'intérêt dans une image synthétique en utilisant le principe du détecteur de Harris et comme score de détection (a)* $score_3$ *et (b)* $score_4$

5 Appariement entre deux images de résolutions différentes

Nous présentons une méthode basée sur l'utilisation des invariants au paramètres intrinsèques de la caméra pour effectuer une mise en correspondance entre images de résolutions différentes prises avec une caméra stationnaire effectuant un zoom.

5.1 Résultat de la détection des points caractéristiques

On applique le détecteur détaillé dans le paragraphe 3 avec nos images. Le détecteur a extrait 422 points dans la première image (voir figure 3(a)) et 433 points dans la

deuxième image (voir figure 3(b)). Le détecteur de Harris repose sur le calcul des courbures locales de la fonction d'auto-corrélation calculée dans une fenêtre d'analyse de taille fixe. En fonction de l'échelle de l'image, les réponses du détecteur sont donc généralement assez différentes. C'est pour cette raison que certains points dans l'image haute résolution n'ont pas été détectés dans l'image basse résolution. Si nous utilisons tous les points pour calculer les invariants alors les deux images ne se superposeront pas (voir figure 3(c)). En effet, certains points sont extraits dans une image et pas dans l'autre. Cependant, grâce au fait que beaucoup de points sont répétés dans les deux images, la reprojection a permis de rapprocher un certain nombre de pairs des points correspondants. Ceci est une conséquence directe de la normalisation des moments du nuage des points.

5.2 Élimination des points non répétés

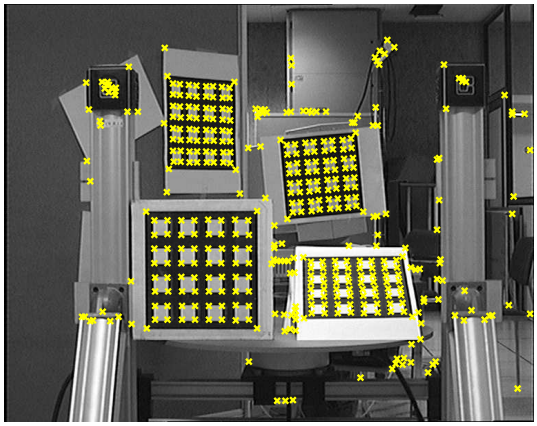
Le but de l'algorithme d'élimination est de supprimer les points qui ne sont pas présents dans les deux images et donc de ne plus en tenir compte lors de l'estimation de la matrice $\mathbf{T}_{p1}\mathbf{T}_{p2}^{-1}$. Deux critères sont utilisés pour l'élimination des points. Premièrement, si les points sont extraits dans les deux images, alors leur reprojection d'une image à l'autre doit être proche. Deuxièmement, un point d'une image reprojeté dans l'autre image doit avoir dans son voisinage au moins un point ayant des propriétés photométriques proches. A chaque point $\mathbf{p}_{ij}$, nous associons 3 descripteurs photométriques:

- $s_1(\mathbf{p}_{ij})$ la moyenne du niveau de gris d'une fenêtre (3×3) centrée en $\mathbf{p}_{ij}$;
- $s_2(\mathbf{p}_{ij})$ le score de la détection défini par (11) ;
- $s_3(\mathbf{p}_{ij})$ un descripteur invariant à l'échelle.

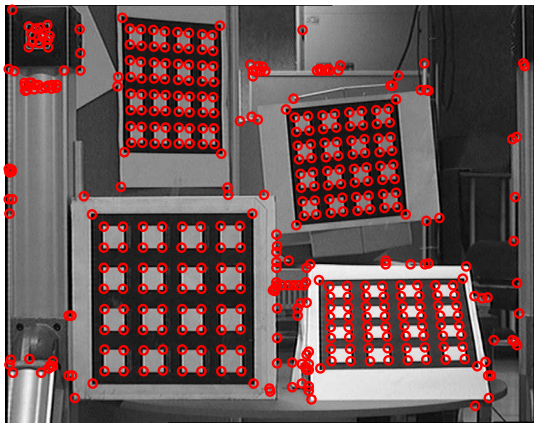
Nous utilisons un cas particulier d'un descripteur invariant à l'échelle proposé dans [12]. Ce descripteur est calculé à l'aide du gradient ∇I et du laplacien ΔI de l'image. Afin d'avoir une fonction continue partout, on définit le descripteur comme suit :

$$
\begin{aligned}
s_3(\mathbf{p}_{ij}) &= 0 \quad \text{si} \quad \| \nabla I(\mathbf{p}_{ij}) \|^2 = 0 \\
s_3(\mathbf{p}_{ij}) &= \frac{\|\nabla I(\mathbf{p}_{ij})\|^2}{|\Delta I(\mathbf{p}_{ij})|} \quad \text{si} \quad \frac{\|\nabla I(\mathbf{p}_{ij})\|^2}{|\Delta I(\mathbf{p}_{ij})|} < 1 \\
s_3(\mathbf{p}_{ij}) &= \frac{|\Delta I(\mathbf{p}_{ij})|}{\|\nabla I(\mathbf{p}_{ij})\|^2} \quad \text{si} \quad \frac{\|\nabla I(\mathbf{p}_{ij})\|^2}{|\Delta I(\mathbf{p}_{ij})|} > 1
\end{aligned}
$$

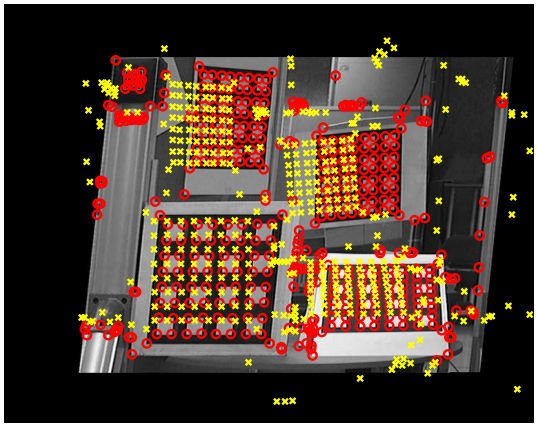
Supposons que le détecteur de Harris a extrait n points $\mathbf{p}_{1k}$, $k \in \{1,...,n\}$ dans l'image 1 et m points $\mathbf{p}_{2j}$, $j \in \{1,...,m\}$ dans l'image 2. Si beaucoup de points sont répétés dans les deux images, les reprojections des points $\mathbf{p}_{2j}$ dans l'image 1 seront proches de leur correspondants comme illustrés dans la figure 3(c). Les points



(a) Image 1



(b) Image 2



(c) Reprojection de l'image 2 sur l'image 1

FIG. 3 – *Points d'intérêt extraits dans les deux images à l'aide du détecteur de Harris.*

isolés ont de faibles probabilités d'avoir des correspondants et peuvent être éliminés comme suit :

Étape 1: Calculer la matrice $\mathbf{T}_{p1}$ en utilisant l'ensemble des points $\mathbf{p}_{1k}$ et la matrice $\mathbf{T}_{p2}$ en utilisant l'ensemble des points $\mathbf{p}_{2j}$. Ensuite, calculer la reprojection des $\mathbf{p}_{2j}$ dans l'image 1 : $\tilde{\mathbf{p}}_{1j} = \mathbf{T}_{p1}\mathbf{T}_{p2}^{-1}\mathbf{p}_{2j}$.

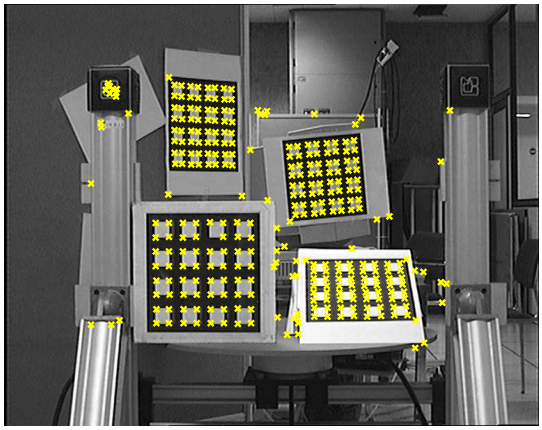
Étape 2. Éliminer $\mathbf{p}_{1k}$ ($\forall k \in \{1,...,n\}$) si $\nexists\, j \in \{1,...,m\}$ tel que :

1. $\| \mathbf{p}_{1k} - \tilde{\mathbf{p}}_{1j} \| < \nu$;
2. $| s_1(\mathbf{p}_{1k}) - s_1(\mathbf{p}_{2j}) | < \tau_1$;
3. $\frac{min\{s_2(\mathbf{p}_{1k}),s_2(\mathbf{p}_{2j})\}}{max\{s_2(\mathbf{p}_{2j}),s_2(\mathbf{p}_{1k})\}} > \tau_2$;
4. $\frac{min\{s_3(\mathbf{p}_{1k}),s_3(\mathbf{p}_{2j})\}}{max\{s_3(\mathbf{p}_{2j}),s_3(\mathbf{p}_{1k})\}} > \tau_3$.

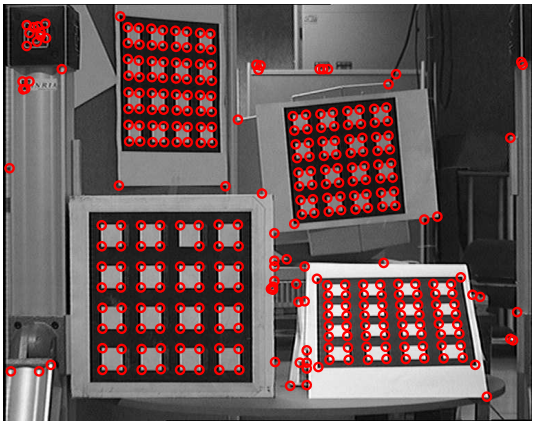
où ν est un seuil variable qui définit la distance maximale autorisée (à la première itération de l'algorithme ν est fixé à ν_{max}), et τ_1, τ_2 et τ_3 sont des seuils choisis tel que $\tau_1 \in [0{,}255]$, $\tau_2 \in [0{,}1]$ et $\tau_3 \in [0{,}1]$. Le seuil τ_1 correspond à la différence maximale autorisée entre la valeur du niveau de gris de deux points correspondants. Les seuils τ_2 et τ_3 permettent de savoir respectivement si deux scores de Harris ou deux descripteurs invariants à l'échelle sont proches ou pas. D'une manière analogue, $\forall j \in \{1,...,m\}$, on élimine $\mathbf{p}_{2j}$ si $\nexists k \in \{1,...,n\}$ tel que les conditions (1), (2), (3) et (4) sont vérifiées. Si au moins un point est éliminé alors reprendre l'algorithme à partir de l'étape 2. Sinon, continuer.

Étape 3: Si $\nu < \nu_{min}$ alors arrêter. Sinon, réduire ν : $\nu = \gamma\nu$ avec $0 < \gamma < 1$. Reprendre l'algorithme à partir de l'étape 1.

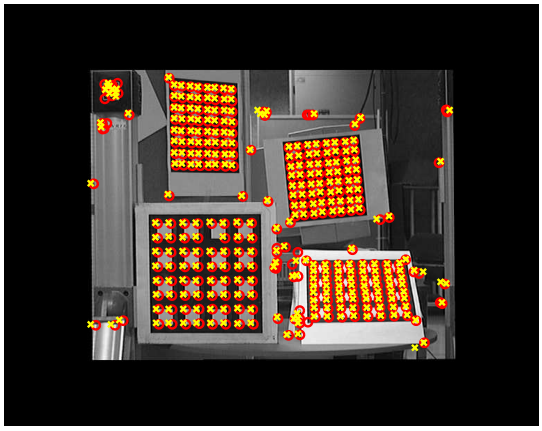
A chaque itération, des points isolés sont éliminés. Par conséquent, on affine l'estimation des matrices $\mathbf{T}_{p1}$ et $\mathbf{T}_{p2}$ et la reprojection des points $\mathbf{p}_{2j}$ dans l'image 1 sont de plus en plus proches de leur correspondants. La valeur des seuils que nous utilisons sont choisis comme suit : $\nu_{max} = 100$, $\nu_{min} = 25$, $\gamma = 0.9$ (c'est-à-dire que nous commençons la recherche des correspondants dans un rayon de 100 pixels, ce rayon décroît d'un facteur de 0.9 à chaque itération et nous arrêtons l'algorithme quand le rayon devient inférieur à 25 pixels), $\tau_1 = 64$, $\tau_2 = \tau_3 = 0.5$ (nous considérons que deux points peuvent être correspondants si la différence de leur valeur de niveau de gris ne dépasse pas 64 et si le rapport de leur score de Harris et celui de leur descripteur invariant à l'échelle sont entre 0.5 et 1). On peut remarquer d'ores et déjà que nous avons une très bonne estimation de la matrice $\mathbf{T}_{p1}\mathbf{T}_{p2}^{-1}$ à l'issue de l'algorithme de l'élimination puisque les croix jaunes et les



(a) Image 1



(b) Image 2



(c) Reprojection de l'image 2 sur l'image 1

FIG. 4 – *Points restants à l'issue de l'algorithme d'élimination.*

cercles rouges de la figure 4(a)(b) coïncident si l'on reprojetée l'image 2 dans l'échelle de l'image 1 (voir figure 4(c)). On remarque également qu'il est possible qu'un certain nombre de points présents dans les deux images et qui sont en correspondance soient éliminés (donc ne sont pas utilisés pour le calculs des invariants). Cependant, ces points seront réintroduits dans l'appariement final.

5.3 Appariement final

A la fin de l'algorithme d'élimination, nous n'utilisons que les points ne présentant pas d'ambiguïté afin d'avoir la meilleure estimation possible de $\mathbf{T}_{p1}$ et de $\mathbf{T}_{p2}$. Ces points vérifient le fait qu'il existe un seul prétendant à leur appariement. Ces deux matrices permettent d'avoir la meilleure reprojection $\tilde{\mathbf{p}}_{1j}$. Enfin, un point $\mathbf{p}_{1k}$ est apparié à $\mathbf{p}_{2j}$ si $\mathbf{p}_{1k}$ est le point le plus proche de $\tilde{\mathbf{p}}_{1j}$. Dans la figure 5(a) et (b), on peut voir le résultat final de l'appariement : les 289 mises en correspondance. Dans la figure 5(c) on peut voir que la deuxième image est parfaitement reprojetée dans l'échelle de la première. Si l'on compare la figure 3(c) et la figure 5(c) que tous les points isolés ont été éliminés. Le bilan de l'appariement est décrit dans le tableau 1. Il y figure le nombre de points extraits par le détecteur, le nombre de points restants à l'issue de l'algorithme d'élimination, le nombre de mises en correspondance réalisées et le nombre de faux appariements. La vérification des appariements à été faite à la main. Nous adopterons la même forme de bilan pour les résultats expérimentaux.

	Image 1	Image 2
Points détectés	422	433
Points restants	314	323
Points appariés	289	
Faux appariements	2	

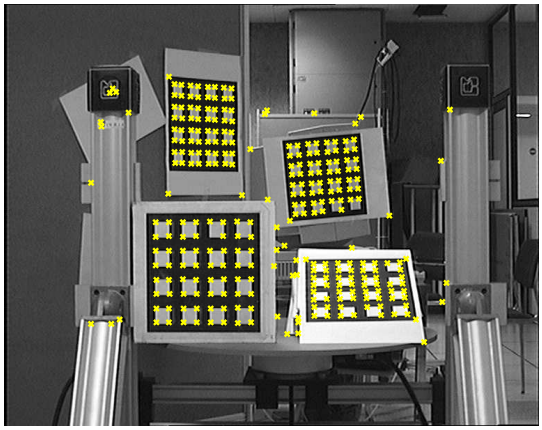
TAB. 1 – *Bilan de l'appariement : le taux d'appariements corrects est d'environ 99%*

6 Résultats expérimentaux

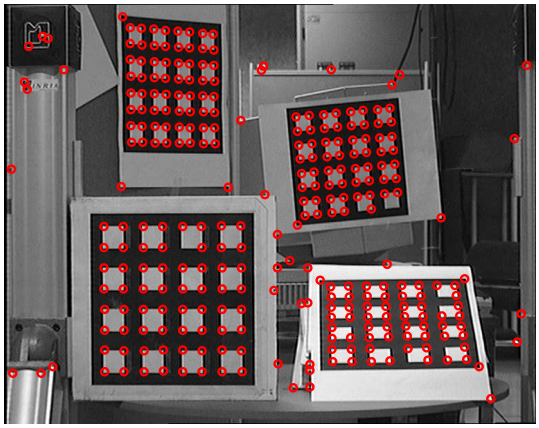
Les exemples présentés dans ce paragraphe s'intéressent à l'appariement entre images d'échelles différentes. Nous traitons les différentes transformations suivantes :

- un zoom pur entre les images ;
- un zoom + de la distorsion.
- un zoom + un faible déplacement de la caméra.

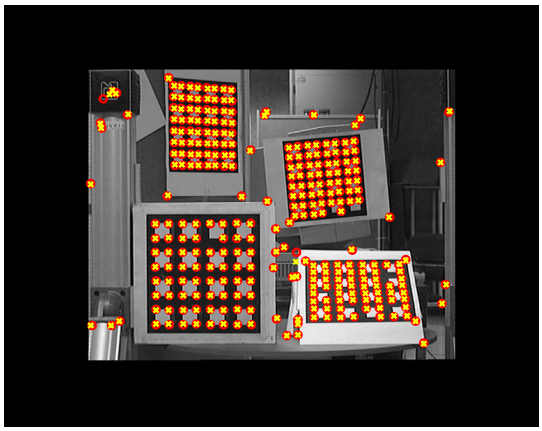
Nous avons essayé de donner des exemples assez variés (scènes d'intérieur / d'extérieur, scènes structurées / non structurées...). Il est important de préciser que la valeur des paramètres utilisés (seuils d'extraction, paramètres



(a) Image 1



(b) Image 2



(c) Reprojection de l'image 2 sur l'image 1

FIG. 5 – *La mise en correspondance finale entre les deux images. Parmi les 289 appariements effectués entre (a) et (b) seulement 2 sont faux.*

de l'algorithme d'élimination...) sont les mêmes pour tous les exemples. Nous n'avons pas eu besoin de modifier ces paramètres pour les adapter à la nature de l'image ou selon la variation de l'échelle entre les deux images. Nous avons représenté les points mis en correspondance par des croix rouges dans le cas où l'appariement est correct et par des triangles jaunes dans le cas d'un faux appariement.

6.1 Cas d'un zoom pur

Nous présentons ici deux exemples d'appariements dans le cas d'un zoom pur entre les deux images. Il faut rappeler tout d'abord que le zoom pur n'est qu'une approximation et qu'une variation de zoom introduit dans la majeure partie des cas une faible variation des paramètres extrinsèques [6]. Ce que nous appelons zoom pur est le changement du zoom sans déplacement de la caméra. Dans le premier exemple (figure 6), la scène est celle qui nous a servi à expliquer l'algorithme. Cependant, cette fois ci, le zoom entre les images est deux fois plus important.

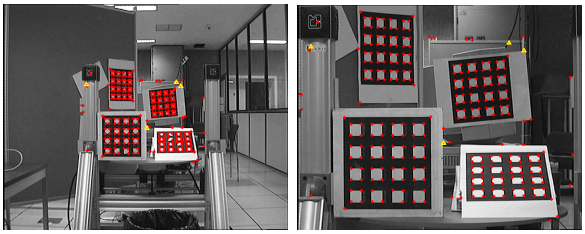


FIG. 6 – *Mise en correspondance entre deux images où le changement d'échelle est de 2.5.*

Nous pouvons voir que la mise en correspondance a bien fonctionné puisque le taux d'appariements corrects est d'environ 99% (voir tableau 2).

	Image 1	Image 2
Points détectés	499	433
Points restants	294	300
Points appariés	284	
Faux appariements	4	

TAB. 2 – *Bilan de l'appariement : le taux d'appariements corrects est d'environ 99%*

Dans le deuxième exemple, il s'agit également d'images d'une scène d'intérieur (figure 7). Dans cet exemple, le changement d'échelle entre les deux images est plus faible que dans le premier exemple mais il y a trois fois moins de points extraits (voir tableau 3).

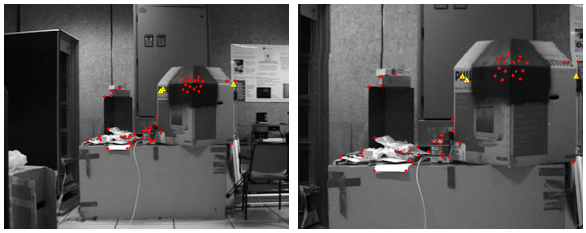


FIG. 7 – *Mise en correspondance entre deux images où le changement d'échelle est de 1.6.*

L'appariement fonctionne bien puisque dans les 50 appariements effectués seulement 3 sont faux.

	Image 1	Image 2
Points détectés	171	132
Points restants	56	66
Points appariés	50	
Faux appariements	3	

TAB. 3 – *Bilan de l'appariement : le taux d'appariements corrects est de 94%*

6.2 Cas d'un zoom + distorsion

La distorsion apparaît dans le cas des focales courtes. C'est une non-linéarité aux bords de la lentille de l'optique employée sur la caméra. Par conséquent, l'approximation du modèle sténopé de la projection perspective n'est plus tout à fait valable [14]. Afin de vérifier l'efficacité de l'algorithme d'appariement proposé, nous utilisons une caméra présentant de la distorsion pour photographier des fleurs dans un vase avec une focale courte et une focale longue (voir figure 8). Le changement d'échelle est de 1.5 entre les deux images.

FIG. 8 – *Mise en correspondance entre deux images où le changement d'échelle est de 1.5.*

Malgré la distorsion de la première image, la mise en correspondance a bien fonctionné avec 109 appariements

dont seulement 5 sont faux (voir tableau 4). Le taux d'appariements corrects est de 95%.

	Image 1	Image 2
Points détectés	227	235
Points restants	123	135
Points appariés	109	
Faux appariements	5	

TAB. 4 – *Bilan de l'appariement : le taux d'appariements corrects est d'environ 95%*

6.3 Cas d'un zoom + faible mouvement de la caméra

Nous nous intéressons ici au cas où la variation d'échelle entre les images est accompagnée d'un faible mouvement de la caméra. Dans ce cas, l'utilisation par les invariants aux paramètres intrinsèques permet de réduire l'effet de la variation d'échelle et l'utilisation des descripteurs photométriques permettent d'effectuer l'appariement. Dans le premier exemple, il s'agit d'une scène d'intérieur prise avec deux focales différentes (voir figure 9) . Une translation et deux rotations entre les deux prises ont été effectuées.

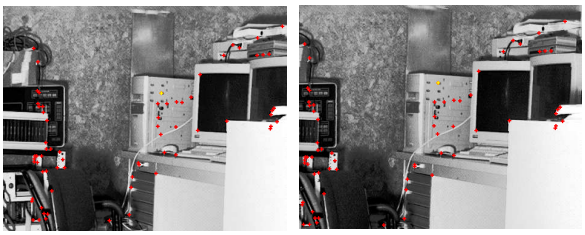


FIG. 9 – *Mise en correspondance entre deux images où le changement d'échelle est de 1.2. Un déplacement de la caméra a été effectué entre les deux images.*

Dans le tableau 5, on peut voir que dans les 69 appariements effectués seulement un appariement est incorrect.

	Image 1	Image 2
Points détectés	111	147
Points restants	61	62
Points appariés	69	
Faux appariements	1	

TAB. 5 – *Bilan de l'appariement : le taux d'appariements corrects est d'environ 99%*

Dans le deuxième exemple, il s'agit d'une scène d'extérieur (voir figure 10) prise avec deux focales différentes. La variation d'échelle entre les images est de 1.4. Un déplacement en translation et en rotation entre les deux prises a été effectué.

FIG. 10 – *Mise en correspondance entre deux images où le changement d'échelle est de 1.4. Un faible déplacement de la caméra a été effectué entre les deux images.*

La mise en correspondance a également bien fonctionné puisque sur les 33 appariements effectués seulement 4 sont incorrects (voir tableau 6).

	Image 1	Image 2
Points détectés	156	186
Points restants	52	47
Points appariés	33	
Faux appariements	4	

TAB. 6 – *Bilan de l'appariement : le taux d'appariements corrects est d'environ 88%*

7 Conclusions et perspectives

L'algorithme de mise en correspondance entre images de résolutions différentes proposé permet d'avoir de très bons résultats dans le cas où un zoom pur est effectué entre les deux images et dans le cas où le zoom est accompagné de la distorsion ou d'un faible déplacement. Les paramètres de l'algorithme n'ont pas été modifiés selon les différents cas. Pour avoir de meilleurs résultats, il est possible d'adapter les paramètres suivant la nature des images et suivant les déplacements effectués entre les images. A cet algorithme, on peut rajouter d'autres couches telles que la corrélation calculée une fois le changement d'échelle a été estimé ou bien l'estimation robuste d'homographie et/ou de matrice fondamentale pour éliminer les faux appariements. Il est possible d'étendre l'approche au cas d'une caméra mobile en couplant l'algorithme présenté dans cet article avec une estimation de déplacement. L'apport de l'algorithme proposé serait alors de réduire le nombre d'inconnus lors

l'estimation puisque l'on s'affranchit de l'estimation de la transformation affine.

Annexe

Considérons la distribution du nuage des points $\mathbf{p}_j$ dans une image donnée et attribuons à tout les points le même poids. Nous montrons ici que la transformation définie par l'équation $\mathbf{q}_j = \mathbf{T}_p^{-1}\mathbf{p}_j = (a_j,b_j,1)$ a pour effet de normaliser les moments d'ordre 2 et d'annuler les moments d'ordre 1 de la distribution des $\mathbf{q}_j$. La matrice $\mathbf{S}_p$ définie dans (4) s'écrit :

$$\mathbf{S}_p = \frac{1}{n}\sum_{j=1}^{n}\mathbf{p}_j\mathbf{p}_j^\top = \frac{1}{n}\sum_{j=1}^{n}\begin{bmatrix} u_j^2 & u_jv_j & u_j \\ u_jv_j & u_j^2 & v_j \\ u_j & v_j & 1 \end{bmatrix}$$

Dans cette matrice, on trouve les moments d'ordre 1 et d'ordre 2 du nuage de points $\mathbf{p}_j$:

$$\mu_{20} = \frac{1}{n}\sum_{j=1}^{n} u_j^2 \qquad \mu_{02} = \frac{1}{n}\sum_{j=1}^{n} v_j^2$$
$$\mu_{10} = \frac{1}{n}\sum_{j=1}^{n} u_j \qquad \mu_{01} = \frac{1}{n}\sum_{j=1}^{n} v_j$$
$$\mu_{11} = \frac{1}{n}\sum_{j=1}^{n} u_jv_j$$

De manière similaire nous pouvons définir une matrice $\mathbf{S}_q$ comme suit :

$$\mathbf{S}_q = \frac{1}{n}\sum_{j=1}^{n}\mathbf{q}_j\mathbf{q}_j^\top = \frac{1}{n}\sum_{j=1}^{n}\begin{bmatrix} a_j^2 & a_jb_j & a_j \\ a_jb_j & b_j^2 & b_j \\ a_j & b_j & 1 \end{bmatrix}$$

Dans cette matrice, on trouve les moments d'ordre 1 et d'ordre 2 du nuage de points $\mathbf{q}_j$:

$$\mu'_{20} = \frac{1}{n}\sum_{j=1}^{n} a_j^2 \qquad \mu'_{02} = \frac{1}{n}\sum_{j=1}^{n} b_j^2$$
$$\mu'_{10} = \frac{1}{n}\sum_{j=1}^{n} a_j \qquad \mu'_{01} = \frac{1}{n}\sum_{j=1}^{n} b_j$$
$$\mu'_{11} = \frac{1}{n}\sum_{j=1}^{n} a_jb_j$$

La transformation a pour effet de modifier les moments d'ordre 1 et d'ordre 2. En effet, la matrice $\mathbf{S}_q$ peut s'écrire sous la forme suivante :

$$\begin{aligned} \mathbf{S}_q &= \frac{1}{n}\sum_{j=1}^{n}\mathbf{T}_p^{-1}\mathbf{p}_j\mathbf{p}_j^\top\mathbf{T}_p^{-\top} \\ &= \mathbf{T}_p^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{p}_j\mathbf{p}_j^\top\right)\mathbf{T}_p^{-\top} \\ &= \mathbf{T}_p^{-1}\mathbf{S}_p\mathbf{T}_p^{-\top} = \mathbf{I} \end{aligned}$$

Par conséquent, par identification, nous avons :

$$\mu'_{20} = \mu'_{02} = 1 \qquad \text{et} \qquad \mu'_{10} = \mu'_{01} = \mu'_{11} = 0$$

Deux moments d'ordre 2 ont été normalisés et un a été annulé, alors que les moments d'ordre 1 ont été annulés. Par exemple, si les points décrivaient une ellipse quelconque dans l'image, ils se transformeraient sur un cercle de rayon unité centré à l'origine.

Références

[1] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 612–618, 2000.

[2] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24(6):381–395, 1981.

[3] W. Forstner. A framework for low-level feature extraction. In *European Conf. on Computer Vision*, pp. 383–394, Stockholm, Sweden, 1994.

[4] B. Hansen and B. Morse. Multiscale image registration using scale trace correlation. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 2202–2208, 1999.

[5] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conf.*, pp. 147–151, 1988.

[6] M. Li and J.-M. Lavest. Some aspects of zoom lens camera calibration. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 18(11):1105–1110, 1996.

[7] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[8] E. Malis. A unified approach to model-based and model-free visual servoing. In *European Conf. on Computer Vision*, pp. 433–447, 2002.

[9] J. Noble. Finding corners. *Image and Vision Computing Journal*, 6(2):121–128, 1988.

[10] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *IEEE Int. Conf. on Computer Vision*, pp. 754–760, 1998.

[11] C. Schmid and R. Mohr. Matching by local invariants. Technical Report 2644, INRIA, 1995.

[12] C. Schmid and R. Mohr. Local grayvalue invariants for image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[13] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *IEEE Int. Conf. on Computer Vision*, pp. 230–235, 1998.

[14] C.C. Slama, C. Theurer, and S.W. Henriksen, editors. *Manual of photogrammetry*. American Society of Photogrammetry, 4th edition, 1980.

[15] A. P. Witkin. Scale-space filtering. In M. A. Fischler and O. Firschein, editors, *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*, pp. 329–332. Kaufmann, CA., 1987.

[16] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Technical Report 2273, INRIA, 1994.