Modèle a contrario pour la mise en correspondance robuste sous contraintes épipolaires et photométriques

Nicolas Noury, Frédéric Sur, Marie-Odile Berger

Modèle a contrario pour la mise en correspondance robuste sous contraintes épipolaires et photométriques

Nicolas Noury Frédéric Sur Marie-Odile Berger

Nancy Université / INPL / INRIA Nancy - Grand Est / LORIA

LORIA, Campus Scientifique - BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France
{noury, sur, berger}@loria.fr

Résumé

La mise en correspondance de points d'intérêt entre deux vues est une des étapes clés en vision par ordinateur, en particulier dans l'analyse de la structure et du mouvement. Après l'extraction de points d'intérêt, deux étapes sont généralement mises en œuvre : la mise en correspondance de ceux-ci en gardant les « meilleurs appariements » selon une mesure de ressemblance photométrique adaptée, puis la sélection des appariements cohérents avec la géométrie induite par le mouvement de la caméra. La présence de motifs répétés, ou des forts changements de point de vue peuvent générer de nombreux appariements aberrants. Nous présentons une méthode a contrario étendant celle de Moisan et Stival [12], qui regroupe ces deux étapes. L'approche proposée ne nécessite pas de paramètre critique et permet un gain significatif en nombre d'appariements obtenus et en précision, en particulier en présence de motifs répétés ou de forts changements des points de vue.

Mots Clef

Mise en correspondance de points d'intérêt, motifs répétés, modèle a contrario.

Abstract

Matching points of interest between two views is one of the keystone of computer vision, especially when considering structure from motion estimation. After extracting points of interest, two steps follow : matching points of interest by keeping only the "best matches" w.r.t. some photometric similarity, then sweep these latter correspondences according to geometric constraints induced by the camera motion. Repeated patterns or widely separated views give lots of false matches. We propose an a contrario approach to these two steps, extending Moisan and Stival's model [12]. The proposed method does not only rid of tricky parameters, but it also shows a significant gain in the number of correspondences, in accuracy (especially when confronted to repeated patterns) and its robustness to wide baseline.

Keywords

Points of interest matching, repeated patterns, a contrario model.

1 Introduction

La mise en correspondance de points d'intérêt extraits de plusieurs images est un des problèmes récurrents de la vision par ordinateur. Les points d'intérêt sont définis à l'aide du gradient de l'image (détecteur de coin de Harris-Stephens) ou de son Laplacien (comme les points Sift [10]). Par « mise en correspondance » il est entendu « appariement de points dans les images qui correspondent au même point de la scène tridimensionnelle ». Il s'agit d'un problème difficile car l'aspect du voisinage de ces points dans les images peut changer fortement lors d'une modification, même limitée, du point de vue. En particulier, cette étape est souvent cruciale dans le problème dit « de la structure et du mouvement ». Le but est alors d'estimer le mouvement relatif d'une caméra entre deux vues d'une même scène tridimensionnelle grâce aux points mis en correspondance d'une image à l'autre, et de reconstruire la position des points de la scène correspondant aux appariements.

Dans cet article, nous présentons une nouvelle méthode statistique pour mettre en correspondance des points d'intérêt entre deux vues. Nous étendons pour cela un modèle précédemment introduit par L. Moisan et B. Stival [12]. Dans la section 1.1 nous expliquons comment se fait généralement l'estimation du mouvement de la caméra entre deux vues, en quoi la mise en correspondance est importante et quels sont les éléments limitants (en particulier présence de motifs répétés et changement de point de vue important). Nous précisons notre contribution dans la section 1.2 et présentons les travaux liés dans la section 1.3. La partie 2 traite du modèle statistique utilisé, et la partie 3 explique l'algorithme proposé et les heuristiques nécessaires. Enfin, nous présentons des résultats expérimentaux dans la partie 4, montrant l'apport de notre méthode pour la mise en correspondance de motifs répétés et dans le cas de changement important du point de vue.

1.1 L'estimation de mouvement

Les algorithmes d'estimation du mouvement d'une caméra à partir de deux vues (images) de la même scène tridimensionnelle débutent classiquement par une mise en corres-

pondance de points d'intérêt :

1. Dans chaque vue, extraction de points d'intérêt et des descripteurs de la photométrie locale associés ;
2. Mise en correspondance des points d'intérêt selon une mesure de ressemblance entre descripteurs ;
3. Filtrage des correspondances précédemment détectées en déterminant l'ensemble le plus compatible avec la géométrie imposée par le mouvement de la caméra.

L'étape 1 a fait l'objet d'une littérature abondante [11], nous utiliserons dans cet article le descripteur Sift de D. Lowe [10] qui est invariant par changement d'échelle et rotation (similitude). Néanmoins notre approche est générale et peut être aisément adaptée à d'autres descripteurs invariants.

L'étape 2 est certainement le point délicat de cette approche : il est très difficile de doter l'espace des descripteurs d'une mesure de ressemblance appropriée. Définir les mises en correspondances en fixant un seuil sur la distance euclidienne entre descripteurs invariants n'est pas réalisable en pratique, ni d'ailleurs en se limitant au plus proche voisin. La méthode la plus répandue dans la littérature, proposée par D. Lowe, est de mettre en correspondance un point d'intérêt de la première vue avec le plus proche voisin dans la seconde vue, à condition que le rapport entre la distance (euclidienne) de ce plus proche voisin et du deuxième plus proche voisin est en dessous d'un certain seuil (bien sûr en dessous de 1). Cette condition permet d'éliminer les appariements ambigus. Néanmoins, il est évident qu'elle élimine également les appariements entre structures répétées pour lequel le rapport est toujours proche de 1. Ce problème apparaît par exemple lors de la comparaison de deux vues de la façade d'un immeuble : une fenêtre particulière devrait être mise en correspondance sur des critères photométriques avec *toutes* les autres fenêtres (pas seulement le plus proche voisin). D'autre part, lorsqu'il y a un (relativement) fort changement de l'aspect de la scène d'une vue à l'autre, la mise en correspondance fournit beaucoup de « faux » appariements, i.e. des appariements entre points des images ne correspondant pas au même point de la scène, car l'invariance par similitude n'est plus suffisante.

Une fois qu'un ensemble de correspondances a été extrait des deux images, l'étape 3 permet d'en sélectionner un sous-ensemble compatible avec un mouvement de caméra réaliste. Dans le cas d'une caméra modélisée selon le modèle de sténopé, les « bons » appariements sont soumis aux contraintes de la géométrie épipolaire, représentée via la matrice fondamentale [8]. Comme l'étape 2 fournit un certain nombre de faux appariements, un choix habituel pour l'étape 3 est d'utiliser une méthode statistique dite *robuste* dérivée de Ransac [6], telles Msac [20] pour sa simplicité, ou d'autres plus élaborées [2, 7]. Ces méthodes nécessitent le délicat réglage d'un certain nombre de paramètres.

Après ces trois étapes, le mouvement de la caméra peut être estimé grâce à l'ensemble des correspondances ainsi trouvé, éventuellement après une nouvelle étape de « mise en correspondance guidée » par cette estimation [8]. Nous n'abordons pas ces derniers aspects ici, mais soulignons que dans tous les cas la liste des points en correspondance à l'issue de l'étape 3 doit être de bonne qualité.

1.2 Contribution

L'algorithme de mise en correspondance générique décrit dans la section précédente pose plusieurs problèmes. Si l'étape 2 (basée uniquement sur une mesure de ressemblance des descripteurs) fournit peu d'appariements, et/ou beaucoup de « faux » appariements, l'étape 3 (qui permet uniquement d'extraire un sous-ensemble consistant avec un mouvement de caméra réaliste) n'a aucune chance de fonctionner.

Notre contribution est de proposer un algorithme et un modèle statistique unifiant les étapes 2 et 3 : on cherche un ensemble d'appariements satisfaisant simultanément la ressemblance entre descripteurs et la cohérence de la position des points d'intérêt avec le déplacement de caméra. Nous nous appuyons sur un modèle probabiliste dit *a contrario* proposé par L. Moisan et B. Stival [12]. Leur algorithme (de type Ransac) présente l'avantage de ne pas nécessiter de paramètre ad hoc. Néanmoins, il nécessite une liste d'appariements pré-établis (donnés par l'étape 2). L'extension à la mise en correspondance sans appariements fournis a priori est seulement esquissée dans [12]. Nous proposons de combiner leur modélisation probabiliste de la contrainte géométrique avec un modèle statistique pour la contrainte photométrique, proposé récemment par J. Rabin, J. Delon, et Y. Gousseau [16, 15]. De plus, ce dernier modèle est basé sur une distance entre descripteurs bien plus performante que la simple distance euclidienne.

Le modèle que nous proposons est particulièrement intéressant en présence de motifs répétés entre images et pour des changements importants d'aspect des images, cas dans lesquels la mise en correspondance au sens des plus proches voisins dans l'étape 2 ne peut pas fonctionner correctement.

Une version préliminaire de ces travaux a été présentée dans [14].

1.3 Travaux liés

L'idée de tenir compte de la ressemblance photométrique pour guider la recherche d'appariements cohérents avec le mouvement de la caméra a fait l'objet de travaux récents, cf [1, 7, 18]. Ces travaux, tous basés sur une adaptation de l'algorithme Ransac, mettent l'accent sur l'accélération de la convergence vers un ensemble de correspondances convenable. L'idée commune est d'utiliser la distance entre descripteurs pour « guider » la recherche de correspondances : l'échantillonnage n'est plus uniforme comme dans Ransac mais s'effectue selon une loi de probabilité reflétant la ressemblance photométrique des correspondances potentielles. Néanmoins, les résultats sont toujours tributaires d'une première recherche de correspondances selon un critère purement photométrique. Le problème posé par les structures répétées n'est pas réellement abordé.

La recherche de correspondances sans appariement photométrique a priori a été, à notre connaissance, pour la première fois proposé par Dellaert et al. [3]. Néanmoins, il s'agit d'une approche purement combinatoire du problème : il est explicitement supposé que le même nombre de points est détecté dans les images, que ces points sont correctement détectés (i.e. qu'ils correspondent tous à un point 3D qui donne naissance à des points d'intérêt dans chaque image, ce qui suppose l'absence d'occultations). Le problème est résolu par maximisation de la vraisemblance géométrique des (très nombreux) appariements possibles, dans une méthode de recuit simulé permettant d'éviter les minima locaux. Si cette approche est intéressante pour contenir la combinatoire importante du problème, les hypothèses semblent trop contraignantes pour qu'elle puisse être utilisée dans un cadre réaliste.

Domke et Aloimonos [5] définissent la vraisemblance des mouvements de caméra en s'appuyant sur un modèle probabiliste mêlant ressemblance photométrique (mesurée par une distance entre descripteurs de Gabor) et cohérence géométrique (distance aux lignes épipolaires). Aucun appariement a priori n'est nécessaire. Néanmoins, la maximisation de la vraisemblance est comme dans [3] un problème combinatoire très difficile (même dans leur contexte où les caméras sont calibrées au préalable, réduisant la représentation du mouvement à la matrice essentielle). La solution présentée par les auteurs permet de trouver une solution lorsque l'on dispose d'une prédiction du mouvement attendu, mais est beaucoup trop coûteuse dans un autre cadre.

2 Un modèle *a contrario* pour l'estimation de la matrice fondamentale sans appariements pré-établis

Dans cette partie, nous présentons une méthode probabiliste de mise en correspondance robuste utilisant la contrainte épipolaire, ainsi que la ressemblance entre descripteurs de points d'intérêt (donnés par Sift).

La modélisation *a contrario*, introduite par A. Desolneux, L. Moisan, et J.-M. Morel [4], consiste à détecter des *groupements* de données élémentaires dont l'existence est très peu probable sous l'hypothèse que ces données sont indépendantes l'une de l'autre. Supposer l'indépendance rend le calcul des probabilités très simple, car la loi jointe est alors le produit des lois marginales qu'il est possible d'estimer empiriquement à partir d'un nombre raisonnable d'observations. Plutôt que de calculer la probabilité d'existence d'un tel groupe, on verra qu'il est plus pratique de dénombrer le nombre de groupes attendus sous hypothèse d'indépendance. Plus cette espérance est faible, moins il est probable que le groupe soit compatible avec l'hypothèse d'indépendance. Le groupe est alors dit *significatif* dans le sens où il est la manifestation d'une causalité sous-jacente. Dans notre cas, un groupe formé d'appariements sera significatif s'il n'est pas explicable par le « hasard » (cas d'appariements indépendants entre eux), mais par une causalité commune (les points d'intérêt sont appariés car ils sont bien les images des mêmes points 3D et sont cohérents avec le mouvement de la caméra).

Donnons quelques notations. Nous supposons disposer de deux vues ($\mathcal{I}_1$ et $\mathcal{I}_2$) de la même scène. Des points d'intérêt (Sift, par exemple) sont extraits de chaque image, ainsi que les descripteurs correspondants. Notons $(x_i, D(x_i))_{1 \leqslant i \leqslant N_1}$ (respectivement $(y_j, D(y_j))_{1 \leqslant j \leqslant N_2}$.) les N_1 (resp. N_2) couples de $\mathcal{I}_1$ (resp. $\mathcal{I}_2$) tels que x_i (resp. y_j) soit les coordonnées d'un point d'intérêt, avec $D(x_i)$ (resp. $D(y_j)$) son descripteur associé. Selon le contexte, nous noterons x_i le point d'intérêt lui-même, ses coordonnées pixels, ou ses coordonnées homogènes dans le plan projectif.

Nous supposons être dans le cadre du modèle de sténopé, et donc pour deux vues dans celui de la géométrie épipolaire. Si on dispose alors de x_i et y_j, les projections dans $\mathcal{I}_1$ et $\mathcal{I}_2$ du même point M de la scène, alors y_j est sur la droite épipolaire associée à x_i dans l'image $\mathcal{I}_2$. Cette droite est représentée par un vecteur normal qui s'exprime Fx_i, où F est la matrice fondamentale. Dans le cas idéal où les descripteurs locaux seraient invariants aux déformations projectives, alors les $D(x_i)$ et $D(y_j)$ sont théoriquement identiques. Cependant, une telle invariance n'existe pas en pratique. Avec Sift, les descripteurs obtenus sont invariants aux changements d'échelle et aux rotations.

Le problème que nous cherchons à résoudre est de trouver un sous-ensemble $\mathcal{S}$ de $\{1, \dots, N_1\} \times \{1, \dots, N_2\}$ et une matrice fondamentale F tels que :

1. La distance entre deux descripteurs est inférieure à un seuil δ_D, ce qui assure que les voisinages des points d'intérêt se ressemblent :
$$\forall (i,j) \in \mathcal{S}, d_D(D(x_i), D(y_j)) \leqslant \delta_D. \qquad (1)$$
2. La distance entre un point de $\mathcal{I}_2$ et la droite épipolaire associée au point en correspondance dans $\mathcal{I}_1$ est inférieure à un seuil δ_G, ce qui permet que la contrainte épipolaire soit satisfaite :
$$\forall (i,j) \in \mathcal{S}, d_G(x_i, Fy_j)) \leqslant \delta_G. \qquad (2)$$

La seconde relation n'est pas symétrique par rapport à $\mathcal{I}_1$ et $\mathcal{I}_2$. Nous ne traitons pas ce point dans cet article.

Nous allons à présent définir le modèle *a contrario* utilisé et les deux (pseudo-)distances d_D et d_G. L'approche statistique proposée fournira des seuils δ_D et δ_G relativement à chaque groupe $\mathcal{S}$.

2.1 Le modèle *a contrario*

On cherche à estimer la probabilité que pour le sous-ensemble $\mathcal{S}$, il existe une matrice fondamentale F pour que les conditions des équations (1) et (2) soient vérifiées avec des seuils δ_D et δ_G. On suppose que la matrice fondamentale F est estimée à partir d'un échantillon minimal s de 7 appariements [19] de $\mathcal{S}$ comme dans l'algorithme Ransac. Il est facile d'estimer cette probabilité sous hypothèse d'indépendance. Plus précisément :

Définition 1. *En supposant que $(x_i, D(x_i))$ et $(y_j, D(y_j))$ sont des variables aléatoires, nous définissons l'hypothèse $\mathcal{H}_0$:*

1. *$(d_D(D(x_i), D(y_j)))_{(i,j)\in\mathcal{S}}$ et $(d_G(x_i, Fy_j))_{(i,j)\in\mathcal{S}}$ sont des variables aléatoires indépendantes.*
2. *$(d_G(x_i, Fy_j))_{(i,j)\in\mathcal{S}\setminus s}$ sont identiquement distribuées et leur fonction de répartition est f_G.*
3. *$(d_D(D(x_i), D(y_j)))_{(i,j)\in\mathcal{S}}$ sont identiquement distribuées et leur fonction de répartition est f_D.*

Nous cherchons alors à estimer :

$$p(\mathcal{S}, F, \delta_G, \delta_D) = Pr\left(\begin{array}{l} \forall (i,j) \in \mathcal{S}, \\ d_D(D(x_i), D(y_j)) \leqslant \delta_D \\ \text{et} \\ d_G(x_i, Fy_j) \leqslant \delta_G \end{array} \middle| \mathcal{H}_0 \right)$$

La probabilité cherchée devient grâce à l'indépendance :

$$p(\mathcal{S}, F, \delta_G, \delta_D) = f_D(\delta_D)^k f_G(\delta_G)^{k-7}$$

où k est le cardinal de $\mathcal{S}$.
Les groupes significatifs seront les groupes pour lesquels cette probabilité est très faible. Néanmoins, on ne peut fixer un seuil uniforme sur cette probabilité car à δ_D et δ_G fixés cela favoriserait automatiquement les grands groupes.
Dans un modèle *a contrario*, on considère plutôt le nombre de fausses alarmes :

Définition 2. *Un ensemble $\mathcal{S}$ d'appariements est ε-significatif s'il existe :*

1. *Deux seuils δ_D et δ_G tels que :*

$$\forall (i,j) \in \mathcal{S}, d_G(x_i, Fy_j) \leqslant \delta_G,$$

$$\forall (i,j) \in \mathcal{S}, d_D(D(x_i), D(y_j)) \leqslant \delta_D.$$

2. *Une matrice fondamentale F calculée sur 7 appariements de $\mathcal{S}$, telle que :*

$$NFA(\mathcal{S}) := 3(\min(N_1, N_2) - 7)k! \binom{N_1}{k} \binom{N_2}{k} \binom{k}{7} \cdot f_D(\delta_D)^k f_G(\delta_G)^{k-7} \leqslant \varepsilon \quad (3)$$

avec k le cardinal de $\mathcal{S}$.

On peut montrer que le Nombre de Fausses Alarmes (NFA) du groupe $\mathcal{S}$ est une majoration de l'espérance du nombre de groupes d'appariements vérifiant la condition 1, sous l'hypothèse $\mathcal{H}_0$. Schématiquement, la probabilité de $\mathcal{S}$ est en effet $p(\mathcal{S}, F, \delta_G, \delta_D)$, que l'on multiplie par le nombre de groupes à tester : il y a $\min(N_1, N_2) - 7$ choix pour $k \geqslant 7$, $\binom{N_1}{k}$ pour les points d'intérêt dans l'image 1, $\binom{N_2}{k}$ choix dans l'image 2, $k!$ choix pour les appariements, $\binom{k}{7}$ choix pour l'échantillon minimal d'appariements servant à estimer F, qui produit 3 matrices fondamentales potentielles (algorithme dit « des 7 points »). Cette définition est esquissée dans [12] (*colored rigidity*).

Un groupe est dit ε-significatif si son NFA est inférieur à ε. Remarquons que le NFA défini dépend en fait des seuils δ_F et δ_G. Les fonctions de répartition f_D et f_G étant croissantes, il est équivalent de définir un groupe ε-significatif à l'aide de la formule (3) où :

$$\delta_D = \max_{(i,j)\in\mathcal{S}} d_D(D(x_i), D(y_j)), \quad (4)$$

$$\delta_G = \max_{(i,j)\in\mathcal{S}} d_G(Fx_i, y_j). \quad (5)$$

2.2 La contrainte géométrique

Nous proposons ici un choix pour d_G et f_G.
Dans [12], il est proposé de définir $d_G(x, Fy)$ comme la distance euclidienne entre x et la droite épipolaire Fy. La fonction f_G est alors :

$$f_G(\delta) = \frac{2D}{A}\delta$$

avec D et A le diamètre et l'aire des images. Sous hypothèse d'indépendance des points d'intérêt et de répartition uniforme dans les images, la probabilité qu'un point choisi au hasard soit à une distance inférieure à δ d'une droite épipolaire est justement majorée par $\frac{2D}{A}\delta$.
Néanmoins, on se rend compte expérimentalement qu'une telle fonction de répartition ne donne pas un poids suffisant à la contrainte géométrique dans l'équation (3). Nous avons donc choisi dans cet article d'utiliser :

$$f_G(\delta) = \left(\frac{2D}{A}\delta\right)^{\alpha}$$

avec α fixé une fois pour toutes à 10 (ce qui s'avère un compromis satisfaisant dans nos vérifications expérimentales).

2.3 La contrainte photométrique

Nous définissons ici d_D et f_D. La plupart des descripteurs sont des histogrammes (circulaires) cumulant les statistiques d'orientation du gradient sur un voisinage des points d'intérêt. Il est connu que la distance euclidienne est mal adaptée à la comparaison d'histogrammes. Il existe d'autres distances plus appropriées que la distance euclidienne pour comparer des histogrammes : dans une certaine mesure la distance du χ^2, et surtout la distance du terrassier EMD (Earth Mover's Distance) [9]. Nous utilisons ici la distance du terrassier circulaire d_{CEMD} proposée par Rabin et al. [16], ainsi que le modèle a contrario associé [15].
Nous ne détaillons pas dans cet article la définition de la distance d_{CEMD} entre histogrammes circulaires. Elle apparaît comme la solution d'un problème de programmation linéaire pour lequel une résolution élégante et efficace est proposée dans [16]. Un descripteur Sift est fait de m histogrammes circulaires ($m = 16$ généralement), et la distance entre deux descripteurs $D^1 = (d_1^1 \dots d_m^1)$ et $D^2 = (d_1^2 \dots d_m^2)$ est définie dans [16] par :

$$\text{dist}(D^1, D^2) = \sum_{i=1}^{m} d_{\text{CEMD}}(d_i^1, d_i^2).$$

Sous hypothèse d'indépendance des histogrammes la probabilité suivante est estimée dans [15] :

$$p(\delta) := \Pr(\text{dist}(D^1, D^2) \leqslant \delta) = \int_0^\delta *_{i=1}^m p_i^{D^1}(u)du$$

avec $*$ le produit de convolution et $p_i^{D^1}$ la fonction de répartition empirique de $d_{\text{CEMD}}(d_i^1, d_i^2)$ lorsque D^2 parcourt l'ensemble des descripteurs de l'image 2.
Avec nos notations, nous définissons :

$$d_D(D(x), D(y)) = p(\text{dist}(D(x), D(y)).$$

Il vient :

$$f_D(\delta) = \Pr(d_D(D(x), D(y)) \leqslant \delta) = \delta.$$

En effet, p est une fonction de répartition ; si on la suppose de plus continue strictement croissante alors :

$$\begin{aligned} f_D(\delta) &= \Pr(p(\text{dist}(D(x), D(y))) \leqslant \delta) \\ &= \Pr(\text{dist}(D(x), D(y)) \leqslant p^{-1}(\delta)) \\ &= p(p^{-1}(\delta)) = \delta. \end{aligned}$$

3 Algorithme

Le but est de trouver le groupe d'appariements présentant le plus petit NFA tel que défini par l'équation (3), avec les distances et fonctions de répartition introduits en 2.2 et 2.3. Une méthode naïve de recherche de ce groupe consiste à tester tous les ensembles possibles d'appariements. Cependant, pour $N = 100$ points d'intérêt extraits dans chaque image, le nombre de groupes à tester est $\sum_{k=1}^{N} k!\binom{N}{k}^2 \simeq 10^{164}$. Faire une recherche par force brute étant irréaliste, nous proposons une heuristique de recherche.

3.1 Diminution de la combinatoire de recherche

La distance d_D issue du modèle *a contrario* de [16, 15] présenté dans la section 2.3 nous permet non seulement de comparer les appariements, mais aussi de sélectionner une liste de *candidats à l'appariement* pour chaque point d'intérêt x de l'image 1 : il ne pourra pas être apparié à tout point de l'image 1, mais seulement aux points y vérifiant :

$$N_1 N_2 d_D(D(x), D(y)) \leqslant \varepsilon$$

La valeur du seuil ε est discutée soigneusement dans [15]. Nous fixons la valeur de ce seuil à $\varepsilon = 10^{-2}$, ce qui permet de garder un nombre assez important de candidats à l'appariement avec un point d'intérêt. Pour des images en 640 par 480 assez proches, en détectant environ 1000 points d'intérêt par image, on obtient 2000 à 4000 candidats à trier.

3.2 Heuristique de recherche par échantillonnage

À ce stade, chaque point d'intérêt x_i de l'image 1 peut être apparié à une liste de N_i points d'intérêt $y_{j_1} \dots y_{j_{N_i}}$ de l'image 2. Il s'agit maintenant de choisir pour chaque x_i un (ou zéro) $y_{j(i)}$ de cette liste. La combinatoire est encore trop élevée pour permettre une recherche exhaustive, nous employons un algorithme par échantillonnage de type Ransac.
Il s'agit d'un algorithme itératif dont la boucle principale contient deux étapes :

1. le tirage de l'échantillon des sept appariements permettant de calculer la matrice fondamentale F,
2. la recherche du groupe le plus significatif formé à partir des appariements potentiels, cohérent avec la matrice F précédente.

Tirage de l'échantillon de sept appariements. Sept points x_i sont tirés au hasard uniformément, et ils sont associés au plus proche voisin au sens de la photométrie $y_{j(i)}$. En effet, comme on le verra dans les expériences, les appariements au sens des plus proches voisins sont souvent corrects. Nous avons donc décidé d'estimer la matrice fondamentale en se basant sur ce critère. Après la coupure proposée dans la section 3.1, il reste peu de candidats à l'appariement pour chaque point x_i. Notre implémentation en retient jusqu'à 10, et on en obtient 2 à 4 en moyenne. La sélection finale des appariements parmis les candidats se fait dans l'étape suivante, une fois que l'on dispose d'une matrice F et donc du critère probabiliste géométrique en plus du critère photométrique.

Sélection de groupes significatifs. On cherche à présent d'autres appariements de manière à former un groupe le plus significatif possible. L'heuristique proposée est la suivante :
1. Pour chaque x_i, sélection de :

$$y_{j(i)} = \underset{y_{j_k}}{\text{argmax}}\{f_D(d_D(D(x_i), D(y_{j_k}))) \cdot f_G(d_G(Fx_i, y_{j_k}))\}$$

et classement croissant des appariements potentiels $(x_i, y_{j(i)})$ selon cette dernière quantité de manière à former une série de groupes d'appariements emboîtés à 7, 8, …, N_1 appariements. C'est ici que l'on sélectionne des appariements qui ne sont pas forcément plus proches voisins, mais n-ième plus proche au sens géométrique.
2. Calcul du NFA de chacun de ces groupes selon l'équation (3) avec les valeurs de δ_D et δ_G données par les équations (4) et (5), et sélection du groupe le plus significatif.
Bien sûr, l'étape 1 ne nous assure pas que le groupe retenu est le plus significatif parmi tous les groupes possibles à matrice F fixée (alors que c'est le cas dans l'algorithme ORSA de [12]). Il s'agit d'une heuristique guidant la recherche basée sur l'observation qu'à k fixé, le groupe le plus significatif sera celui minimisant le produit $f_D(\delta_D) f_G(\delta_G)$.
D'autres heuristiques de recherche dans un contexte similaire sont décrites dans [21].

4 Résultats

Nous montrons d'abord sur des séquences contenant des motifs répétés que notre méthode permet d'obtenir plus d'appariements qu'un algorithme d'appariement classique. Une partie significative des points appariés ne vient pas des descripteurs les plus ressemblants, mais de ceux dont le rang est plus éloigné (n-ième plus proche voisin) selon la distance entre descripteurs. Nous nous comparons pour cela avec une approche classique utilisant les étapes 2 et 3 présentées en section 1.1 :

– appariement d'un point d'intérêt de la première vue avec le plus proche voisin dans la seconde vue, à condition que le rapport entre la distance (euclidienne) de ce plus proche voisin et du deuxième plus proche voisin est en dessous d'un certain seuil (0.6 ici).
– Sélection robuste avec l'algorithme Orsa [12].

L'utilisation successive de ces deux parties est nommée R+O ; notre méthode AC pour *a contrario*.

Les exemples retenus permettent de considérer des situations variées : d'abord, deux exemples, Motif et Nappe (figures 1 et 2), présentants des textures difficiles à apparier, ensuite deux séquences d'intérieur, Rue et Bureau (figures 4 et 5), et une séquence synthétique (Singe, figure 6) présentant une profondeur suffisante permettant l'estimation de la géométrie épipolaire.

Outre les gains en nombre d'appariements retenus, le gain en précision est quantifié dans chaque expérience en mesurant la moyenne des distances d'un point à la droite épipolaire de son correspondant, et en utilisant pour cela la matrice fondamentale F ré-estimée aux moindres carrés sur l'ensemble des appariements retenus.

Dans toutes les images présentées, les appariements corrects sont en bleu, la croix montrant la position du point d'intérêt sur l'image courante, et le segment représentant le mouvement apparent avec la seconde vue. Les appariements non retenus sont indiqués en rouge pour R+O. Pour notre méthode AC, nous ne représentons que les appariements obtenus, l'ensemble des candidats potentiels étant trop vaste.

4.1 Motifs répétés

On considère ici les figures 1 et 2, Motif et Nappe. Dans les deux cas, le mouvement de la caméra est limité. On ne montre les résultats que sur la première des deux images. En présence de motifs répétés, le nombre d'appariements sélectionnés par R+O chute, comme sur la face avant du cube dans la figure 1 où beaucoup de points à l'apparence analogue sont rejetés par le seuil sur le rapport des distances ; 349 appariements sont extraits avec AC contre 179 avec R+O. Sur la figure 2, on détecte de même 697 appariements avec AC contre 501 par R+O. Certains appariements sélectionnés (par exemple, le point en bleu foncé détecté le plus à gauche dans le bas de la figure 2 pour AC), le sont par erreur car ils sont proches de la droite épipolaire associée à leur correspondant : il n'est pas possible de les filtrer dans ce cas, ils sont en effet à la fois cohérents pour

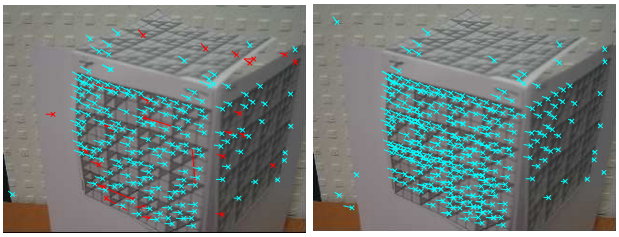

FIG. 1 – Motif - A gauche, R+O ; A droite, AC. La scène présente des motifs répétés : notre méthode permet d'obtenir plus de points détectés dans les zones contenants de tels motifs, où l'ambiguïté pour réaliser la mise en correspondance est importante.

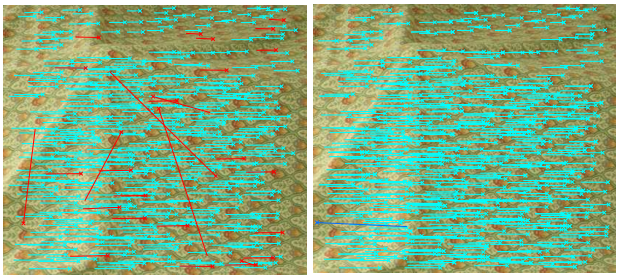

FIG. 2 – Nappe [17] - A gauche, R+O ; A droite, AC. Notre méthode permet d'extraire ici aussi plus d'appariements.

les contraintes géométriques et photométriques. Ici la distance entre un point et la droite épipolaire associée à son correspondant est similaire pour les deux méthodes (0,11 pixel). Ces nombreux points d'intérêt supplémentaires correspondent à des points qui ne sont pas les plus proches voisins pour la distance entre descripteurs, comme indiqué sur le tableau 1. La figure 3 présente les appariements retenus à partir du second rang (n-ième plus proche voisin avec $n > 1$) pour la distance entre descripteurs. Ils se situent bien sur les textures répétitives. Ceci montre que notre méthode n'est pas réductible à remplacer uniquement la distance euclidienne par la distance EMD dans la méthode R+O.

	Nombre d'appariements								
Rang	1	2	3	4	5	6	7	8	9
Motif	299	20	12	8	5	1	2	1	0
Singe 1	118	22	28	8	14	17	13	9	4
Rue	397	18	4	7	0	0	0	0	0

TAB. 1 – Nombre d'occurrence des n-ièmes plus proches voisins sélectionnés par la méthode AC. L'utilisation des appariements de rang supérieur à 2 apporte d'autant plus de nouvelles correspondances que la scène présente des motifs répétés.

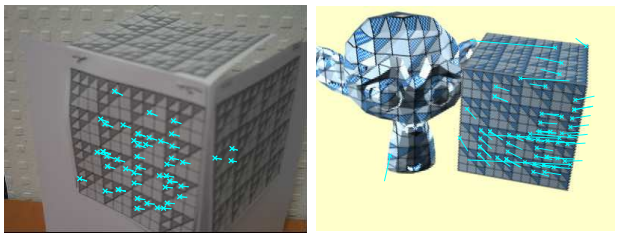

FIG. 3 – Détails de la sélection par AC pour les figures Motif et Singe 1. Sont uniquement affichés les appariements à partir du rang 2 pour la ressemblance entre descripteurs.

4.2 Estimation de mouvement

Maintenant que l'on a montré que notre méthode se comporte bien sur des scènes composées essentiellement de motifs répétés, il est nécessaire de valider qu'elle permet aussi la sélection des appariements et le calcul de la matrice fondamentale pour des images où le mouvement est plus important, comme dans celles rencontrées habituellement lorsque l'on se déplace à l'intérieur d'un bâtiment. C'est le cas de la figure 4 qui présente des points d'intérêt difficiles à apparier (motifs répétés sur la moquette, texture difficile sur les murs). Nous obtenons ici, outre un nombre plus important de correspondances, un gain en précision d'estimation du mouvement, la distance moyenne d'un point à la droite épipolaire associée à son correspondant étant de 0,42 pixel contre 0,82 autrement. Les gains en nombre d'appariements et en précision (distance à l'épipolaire), ainsi que le nombre total de candidats, après R pour R+O, après diminution de la combinatoire pour AC (cf. section 3.1), sont résumés dans le tableau suivant :

	R+O			AC		
	total	**ok**	**préc.**	**total**	**ok**	**préc.**
Motif	206	179	0.13	2177	349	0.11
Nappe	501	472	0.11	4368	697	0.11
Rue	307	277	0.82	1930	409	0.42
Bureau	42	34	1.01	269	79	0.69

TAB. 2 – Comparaison des deux algorithmes : nombre d'appariements total, retenus (ok), et distance moyenne des points appariés à la droite épipolaire correspondante en pixels (préc.) pour les correspondances retenues.

Lorsque le mouvement apparent est fort (Bureau, figure 5), notre méthode permet aussi d'obtenir plus de mises en correspondance correctes.
Des points d'intérêt supplémentaires pourraient être obtenus en modifiant le seuil dans la méthode R+O, puisque les points supplémentaires sont des plus proches voisins pour la distance photométrique. L'avantage de notre méthode ici est ne pas nécessiter ce genre de réglage. Par exemple, sur la figure 4, Rue, avec un seuil sur le rapport de 0.8, plutôt que celui de 0.6 choisi par D. Lowe, on obtient avec R+O 345 appariements sélectionnés parmi 473 et une précision de 0,51 pixel ; sur Bureau (figure 5), on a alors 72 appariements sur 125 et une précision de 0,92 pixel (à comparer avec les résultats du tableau 2). On note quelques appariements aberrants non détectables car compatibles avec la contrainte géométrique, indiqués en bleu foncé : sur le pied de la table dans Bureau, sur le bord gauche du mur foncé et en haut à droite sur le mur du fond dans Rue.

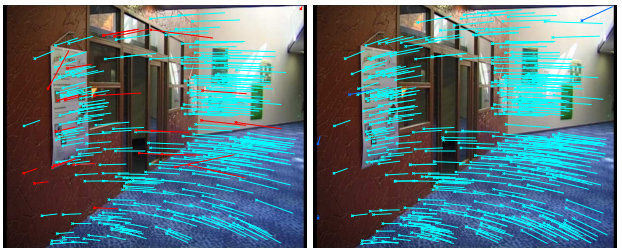

FIG. 4 – Rue - A gauche, R+O ; A droite, AC. On note la présence de quelques appariements aberrants mal filtrés : en utilisant la contrainte épipolaire, ceux-ci sont impossibles à retirer car compatibles avec la géométrie trouvée.

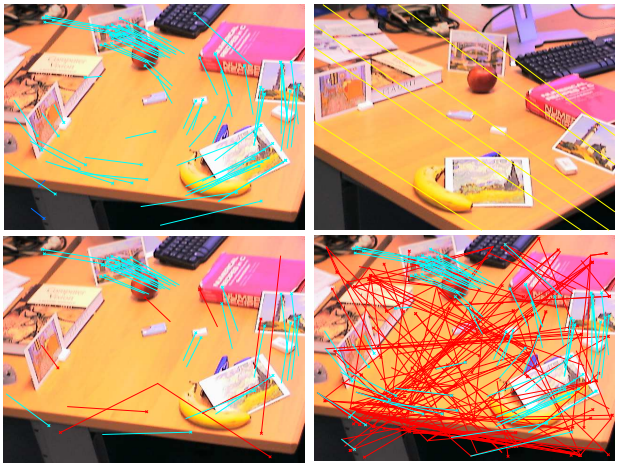

FIG. 5 – Bureau - En haut à gauche, les appariements extraits avec AC, à droite, l'autre vue ainsi que quelques droites épipolaires. Noter le fort mouvement apparent. En bas à gauche, les point d'intérêt mis en correspondance via R+O. En bas à droite, la totalité des candidats à la mise en correspondance pour AC. Le nombre d'appariements testés par AC est bien supérieur à celui de R+O.

Certains appariements qui ne correspondent pourtant pas à des motifs répétés sont retenus par la méthode AC. Ces points correspondent soit aux appariements de rang 1 pour la distance photométrique rejetés par R+O, soit à des appariements de rangs supérieurs sélectionnés grâce à la contrainte géométrique.
Nous sommes cependant limités par l'invariance des descripteurs de points d'intérêt extraits par Sift, qui ne permet pas d'avoir des candidats potentiels lorsque le changement d'apparence est trop important, en particulier dans la figure 5, ou lorsque certaines parties de la scène sont trop déformées comme la face supérieure du cube sur les figures 1 et 6.
En ce qui concerne le temps de calcul, la méthode AC est plus coûteuse que la méthode R+O. Le rapport de vitesse est de 1 à 10, soit une minute pour des images en 640 par 480. Le calcul de probabilités empiriques de ressemblance

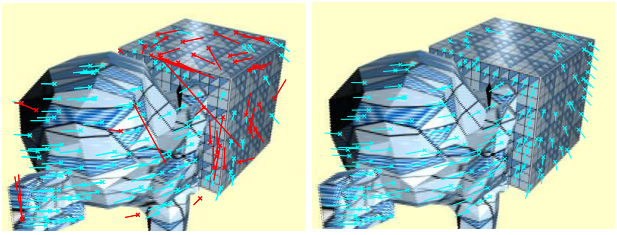

FIG. 6 – Singe 2 - A gauche, R+O ; A droite, AC. 172 appariements extraits avec AC contre 96 pour R+O. Peu d'appariements sont extraits sur la face supérieure du cube.

entre descripteurs est lui bien plus long qu'une mise en correspondance via le critère du rapport de distance, et représente la quasi totalité de la différence en temps de calcul. L'ensemble considéré d'appariements candidats pour le filtrage robuste est lui aussi plus important, d'où un coût de recherche aussi plus élevé.

5 Conclusion

Nous avons présenté dans cet article une méthode *a contrario* qui permet un appariement robuste de points d'intérêt, en prenant en compte simultanément les contraintes géométriques et photométriques. L'apport de cette approche, outre qu'elle ne nécessite pas de paramètre à modifier pour chaque expérience, est un gain significatif en nombre d'appariements obtenus et en précision, en particulier en présence de motifs répétés, ainsi que sa robustesse à des forts changements de point de vue.

Des situations posent encore des difficultés pour résoudre les ambiguïtés de mise en correspondance des motifs répétitifs. En particulier, quand la direction des alignements de motifs répétés est proche de celle des droites épipolaires induites par le mouvement. Dans un tel cas, l'utilisation d'une approche sur trois vues, permettant de durcir la contrainte géométrique (la correspondance point / droite devient une correspondance point / intersection de deux droites), devrait permettre d'apparier ces points.

Cependant, dans le cas de fortes variations d'apparence, notre approche est naturellement limitée par le degré d'invariance du descripteur local utilisé. Une stratégie du type de celle de A-Sift [13], permettant une invariance affine complète, semble prometteuse.

Références

[1] O. Chum and J. Matas. Matching with PROSAC - Progressive Sample Consensus. In *CVPR*, volume 1, pages 220– 226, 2005.

[2] O. Chum, J. Matas, and J. Kittler. Locally Optimized RANSAC. In *DAGM*, 2003.

[3] F. Dellaert, S.M. Seitz, C. Thorpe, and S. Thrun. EM, MCMC, and chain flipping for structure from motion with unknown correspondence. *Machine Learning*, 50(1-2), 2003.

[4] A. Desolneux, L. Moisan, and J.-M. Morel. Maximal Meaningful Events and Applications to Image Analysis. *Annals of Statistics*, 31(6), 2003.

[5] J. Domke and Y. Aloimonos. A Probabilistic Notion of Correspondence and the Epipolar Constraint. In *3DPVT*, 2006.

[6] M. Fischler and R. Bolles. Random Sample Consensus : A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6) :381–395, 1981.

[7] L. Goshen and I. Shimshoni. Balanced Exploration and Exploitation Model Search for Efficient Epipolar Geometry Estimation. In *ECCV*, volume 3952, pages 151–164, 2006.

[8] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN : 0521623049, 2000.

[9] H. Ling and K. Okada. Diffusion distance for histogram comparison. In *CVPR*, volume 1, pages 246–253, 2006.

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2) :91–110, 2004.

[11] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *CVPR*, 2003.

[12] L. Moisan and B. Stival. A Probabilistic Criterion to Detect Rigid Point Matches between Two Images and Estimate the Fundamental Matrix. *IJCV*, 57(3) :201–218, 2004.

[13] J.M. Morel and G. Yu. ASIFT : A New Framework for Fully Affine Invariant Image Comparison. *SIAM Journal on Imaging Sciences*, 2(2) :438–469, 2009.

[14] N. Noury, F. Sur, and M.-O. Berger. Fundamental matrix estimation without prior match. In *ICIP*, 2007.

[15] J. Rabin, J. Delon, and Y. Gousseau. A contrario matching of SIFT-like descriptors. In *ICPR*, 2008.

[16] J. Rabin, J. Delon, and Y. Gousseau. Circular Earth Mover's Distance for the comparison of local features. In *ICPR*, 2008.

[17] D. Scharstein and R. Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *IJCV*, 47 :7–42, 2001.

[18] B. Tordoff and D. W. Murray. Guided Sampling and Consensus for Motion Estimation. In *ECCV*, volume 2350, pages 82 – 98, 2002.

[19] P. Torr and D. Murray. Outlier Detection and Motion Segmentation. In P. S. Schenker, editor, *Sensor Fusion VI*, pages 432–443. SPIE volume 2059, 1993.

[20] P. Torr and A. Zisserman. MLESAC : A new robust estimator with application to estimating image geometry. *CVIU*, 78 :138–156, 2000.

[21] W. Zhang and J. Kosecka. Generalized RANSAC Framework for Relaxed Correspondence Problems. In *3DPVT*, pages 854–860, 2006.