Dix-septième colloque GRETSI, Vannes, 13-17 septembre 1999

Mise en correspondance multi-échelles

Nicolas Allezard, Michel Dhome, Frédéric Jurie

Laboratoire des Sciences et Matériaux pour l'Electronique, et d'Automatique (LASMEA)
UMR 6602 du CNRS
Université Blaise-Pascal de Clermont-Ferrand
F-63177 AUBIERE Cedex, France
allezard@lasmea.univ-bpclermont.fr, dhome@lasmea.univ-bpclermont.fr,
jurie@lasmea.univ-bpclermont.fr

Résumé – Cette article traite de la mise en correspondance de primitives de type point d'intérêt. Ces primitives sont extraites automatiquement de zones de l'image riches en information de luminance. Elles sont donc particulièrement intéressantes dans un contexte de mise en correspondance. Nous proposons ici une description locale et multi-échelles de ces primitives directement basée sur le signal de luminance. Cette description est invariante aux rotations et translations de l'image et très robuste aux changements d'échelles. L'algorithme de mise en correspondance comprend 4 étapes : l'extraction des primitives, leurs caractérisation, la recherche des points semblables, et enfin le filtrage des couples erronés.

Abstract – This article deals with keypoint features matching. These points are located in image zones having a high information content. They are thus particularly interesting in a matching context. We propose a local and multi-scale description of such features directly based on the luminance signal. This description is insensitive to image rotation and translation and furthermore has also a very good robustness to scale changes.The matching algorithm includes 4 main stages: feature extraction, feature description, features matching, and at last filtering of erroneous pairs.

1 Introduction

La reconnaissance automatique d'objets reste une grande problématique en vision artificielle. L'outil fondamental d'un tel processus est la mise en correspondance des primitives de l'image étudiée avec celles présentes dans la base de données. L'utilisation de la texture des objets semble être une voie prometteuse pour la reconnaissance d'objets complexes. Ainsi, un grand nombre d'approches directement basées sur le signal de luminance ont été proposées ces dernières années. Dans cet article, nous proposons un algorithme de mise en correspondance multi-échelles basé sur une description locale de ces primitives. Cette description est invariante aux rotations et translations de l'image et très robuste aux changements d'échelles.

2 Approche proposée et travaux similaires

La méthode que nous proposons suit un paradigme classique en vision artificielle : l'extraction des primitives, leur caractérisation, l'étape d'appariement, le filtrage des couples erronés.

La première étape consiste en l'extraction des points d'intérêt des images. Ces points sont situés dans des zones de l'image riches en information de luminance. Ils sont donc particulièrement intéressants dans un contexte de mise en correspondance. Ces points sont extraits automatiquement des images à l'aide du détecteur de coins proposé par Harris et Stephens [2] et amélioré par Schmid [5]. Ce détecteur a été choisi en raison de sa bonne répétabilité. C'est à dire pour sa capacité à extraire dans des images représentant la même scène, mais acquises dans des conditions différentes (luminosité, zoom, mouvement de la caméra), des points caractéristiques correspondant à des mêmes entités 3D.

Les points d'intérêt sont ensuite caractérisés par un vecteur directement basé sur le signal de luminance. Ce vecteur est invariant pour les rotations et translations planaires de l'images, et assez robuste aux transformations projectives. Nous proposons de plus, une implémentation multi-échelles de la description locale des points d'intérêt de façon à mettre en correspondance les primitives dans le cas d'importants changements d'échelles (la taille de l'image peut être divisée ou multipliée par deux).

La mesure de la similarité des vecteurs est réalisée par le calcul d'un score de corrélation centré et normalisé permettant la prise en compte d'un changement affine de la luminosité entre images. La complexité de la mise en correspondance est réduite par l'utilisation d'une méthode pyramidale.

Enfin, deux contraintes géométrique simples sont utilisées pour rejeter les faux appariements.

Les deux problèmes majeurs dans un algorithme d'appariement sont : la robustesse de la représentation des primitives, le coût algorithmique de la mise en correspondance. Un grand nombre de solutions ont été proposées afin de résoudre l'un ou les deux de ces problèmes.

Nous citerons ici le travail réalisé par Schmid dans sa thèse ayant pour sujet l'indexation d'images. La descrip-

tion des points d'intérêt est ici réalisée grâce à l'utilisation d'invariants différentiels [3]. Après une extraction des points d'intérêt réalisée par l'algorithme d'Harris et Stephens, les primitives sont caractérisées par un vecteur invariant aux rotations et translations de l'image. Une approche multi-échelles est également développée. La mesure de la similarité des points est effectuée par le calcul de la distance de Mahalanobis entre les vecteurs. La matrice de covariance des vecteurs doit donc être estimée. Cette étape d'apprentissage est réalisée par le suivi de points d'intérêt d'images acquises sous différentes conditions. Cette étape n'est pas un problème dans un contexte d'indexation d'images. Mais, dans le cas de l'appariement de primitives extraites d'images inconnues, la matrice de covariance peut être mal estimée et donc conduire à de mauvais résultats.

3 Caractérisation des points d'intérêt

Le vecteur attaché à un point d'intérêt est directement basé sur le signal de luminance voisin. Ce vecteur rassemble des échantillons de niveaux de gris acquis sur des cercles centrés sur le point d'intérêt. Cet échantillonnage est réalisé de façon pyramidale. Ainsi, les échantillons ne sont pas acquis dans l'image originale mais dans N images lissées (N étant le nombre de cercles d'échantillonnage). La quantité de lissage relative à chaque cercle augmente avec le rayon de ces derniers. La fonction de lissage utilisée est celle donnée par la primitive de l'opérateur optimal de dérivé proposé par Deriche [1]. Cette fonction possède de bonnes capacités de filtrage et peut être facilement implémentée récursivement.

Les deux avantages de cette représentation des points d'intérêt sont les suivants.

Premièrement, elle permet de décrire le voisinage d'un point avec un vecteur peu important et donc, de réduire la complexité de mise en correspondance.

Deuxièmement, l'appariement est plus robuste aux changements de points de vue. En effet, lors d'une transformation projective, les plus importantes perturbations se trouvent sur les pixels les plus éloignés du point d'intérêt. Le fort lissage appliqué sur ceux-ci permet d'accroître la stabilité de la représentation.

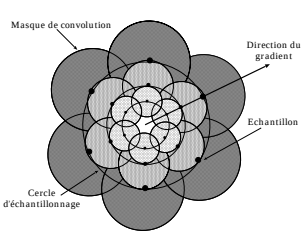

FIG. 1: *Echantillonnage invariant en rotation du voisinage d'un point d'intérêt*

Dans le cas d'une rotation de l'image, l'échantillonnage circulaire doit être décalé d'un angle localement stable par rapport au motif voisin au point d'intérêt. Cet angle peut être celui formé par la direction du gradient au point d'intérêt avec les lignes de l'image (voir figure 1). Cependant, cette direction est très sensible au bruit ainsi qu'à la précision de détection des points d'intérêt. Pour cette raison, nous calculons une direction "moyenne" du gradient sur une fenêtre circulaire centrée sur le point d'intérêt afin d'obtenir une direction stable.

4 Caractérisation multi-échelles

Dans le cas d'un fort changement d'échelle, la méthode de caractérisation des primitives présentée au paragraphe 3 doit être modifiée.

Considérons deux images I_1 et I_2, où I_2 a subi un changement d'échelle Δ. Pour la fonction $f(\overrightarrow{x})$, qui décrit le signal de luminance bi-dimensionnel au voisinage d'un point, le facteur d'échelle Δ peut être représenté par un simple changement de variable : $f(\overrightarrow{x})=g(u(\overrightarrow{x}))$ où $u(\overrightarrow{x}) = \Delta\overrightarrow{x}$. Donc, le rayon de chaque cercle d'échantillonnage ainsi que les tailles des masque de convolution doivent être modifiés relativement au facteur d'échelle. De plus, la direction du gradient est également sensible au changement d'échelle. La taille de l'opérateur utilisé pour calculer les dérivées en x et y doit donc être changée en conséquence.

La modification de l'échantillonnage n'est cependant pas suffisante dans le cas d'un fort changement d'échelle. En effet, les résultats de la mise en correspondance ne sont pas limités par la méthode de caractérisation locale, mais par l'algorithme employé pour l'extraction des points d'intérêt.

L'algorithme développé par Harris et Stephens est construit sur une idée similaire de celui proposé par Moravec [4]. Harris et Stephens estiment l'autocorrélation locale du signal de luminance à partir ses dérivées en x et y. Dans sa thèse, Schmid montre qu'une extraction multi-échelles des points est possible grâce à l'adaptation de la taille de l'opérateur de dérivation : dans le cas d'un facteur d'échelle Δ entre images, cette taille doit être multipliée par Δ.

Dans le cas où le facteur d'échelle entre images est inconnu, ces deux étapes doivent être réalisées à différentes échelles dans une des deux images. Les expériences présentées au paragraphe 7 montrent que cinq échelles sont suffisantes $(0.5, 1/\sqrt{2}, 1.0, \sqrt{2}, 2.0)$ afin d'obtenir de bons résultats pour des variations du facteur d'échelle comprises entre 0.5 et 2.

5 Recherche des points correspondants

La mesure de la similarité des points est réalisée par le calcul d'un score de corrélation centré et normalisé :

$$S = \frac{\sum_{i=1}^{n}(V_{1i} - \overline{V_1})(V_{2i} - \overline{V_2})}{\sqrt{\sum_{i=1}^{n} {V_1}_i^2 \times \sum_{i=1}^{n} {V_2}_i^2}}$$

V_1 et V_2 sont les deux vecteurs de caractérisation, $\overline{V_1}$ et $\overline{V_2}$ leurs moyennes respectives.

Bien que plus long à calculer qu'un score classique, ce score permet de prendre en compte un changement affine de luminosité entre images. Deux points sont associés si leur score de corrélation est supérieur au seuil de rejet. La valeur du seuil a été déterminée expérimentalement, et fixée à 0.95.

De manière à réduire le temps de mise en correspondance, une recherche pyramidale sur les vecteurs a été mise en place. Cette recherche tire partie de la méthode particulière d'échantillonnage du signal voisin aux points d'intérêt. L'idée de base consiste à comparer les plus basses fréquences du motif autour du point, garder les meilleurs appariements, puis introduire plus de détails et comparer les points restants.

Ainsi, le score de corrélation est premièrement calculé sur les échantillons du dernier cercle (le plus éloigné du point) qui représentent la plus grande zone image. Les plus hauts scores sont retenus, puis le score est calculé sur les deux derniers cercles. Le processus est répété jusqu'au cercle le plus proche. Dans les expériences présentées au paragraphe 7, les nombres de vecteurs retenus à chaque étape sont les suivants : 20, 10, 8, 5, 2 et 1.

6 Contrainte géométrique

Afin d'améliorer les résultats, deux filtrages des couples erronés sont mis en oeuvre. Pour le premier, nous recherchons à quelle échelle l'extraction et la caractérisation des points de la première image donnent lieu au plus grand nombre d'appariements. Les points qui non pas été extraits à cette échelle ou à celles immédiatement inférieure ou supérieure sont rejetés.

Le second filtrage est basé sur une contrainte géométrique semi-locale. Il consiste à ne retenir que les couples de points ayant au moins 50% de leur n plus proches voisins eux-mêmes mis en correspondance.

7 Résultats expérimentaux

Dans cette partie, sont présentées quelques expériences réalisées afin de valider notre approche.

La taille du vecteur est ici de 36 composantes. Pour cela, nous avons utilisé 6 cercles d'échantillonnage et 6 échantillons par cercle. Ces derniers ont pour rayon 1.0, 2.0, 4.0, 8.0, 12.0, 18.0 pixels. Les valeurs de α (fixant la taille des opérateurs de lissage) correspondantes sont : 6.0, 3.0, 1.5, 0.75, 0.5, 1/3.

Le grand nombre de points ne permet pas une vérification manuelle des résultats. La validation des appariements est donc effectuée par le calcul de l'homographie représentant le mouvement de la caméra. L'homographie est estimée grâce à une méthode robuste utilisant un filtre médian et tolérant jusqu'à 50% de couples erronés. Cependant, l'utilisation d'une homographie limite le type de scènes observées à celles pratiquement planes.

Dans un contexte multi-échelles, le temps de mise en correspondance est approximativement de 3 secondes pour 4000 et 800 points extraits extraits de la première image et deuxième image. Ce temps peut être grandement diminué par l'utilisation de techniques d'indexation de base de données.

7.1 Rotations entre images

Ici sont présentés les résultats obtenus dans le cas d'une rotation de l'image. Les premières et dernières images de la séquence sont présentés ci-dessous (figure 2). Le pas entre images est d'aproximativement de 60^o.

FIG. 2: *Première et dernière image de la séquence Rotation.*

TAB. 1: Résultats de la séquence Rotation.

Image	Nombre d'appariements	Couples corrects	Pourcentage
1	175	175	100.0
2	155	155	100.0
3	164	164	100.0
4	197	197	100.0
5	173	172	99.4
6	223	223	100.0

7.2 Variation de la luminosité

Cette série de tests illustre la robustesse de la méthode face à un changement de luminosité entre images.

FIG. 3: *Première et dernière image de la séquence Luminosité.*

TAB. 2: Résultats de la séquence Luminosité.

Image	Nombre d'appariements	Couples corrects	Pourcentage
1	18	18	100.0
2	101	101	100.0
3	221	221	100.0
4	177	177	100.0
5	66	66	100.0
6	39	38	97.4

Bien que le nombre de points extraits de la deuxième

image diminue fortement si l'image est très sombre, le pourcentage de couples corrects reste très satisfaisant.

7.3 Mise en correspondance multi-échelles

Cette série d'images a été obtenues en changeant la focale de la caméra. Les facteurs d'échelles indiqués dans les résultats ont été estimés grâce à la matrice d'homographie.


Image de référence

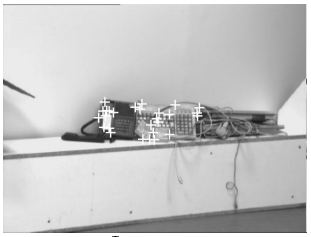
Image 1

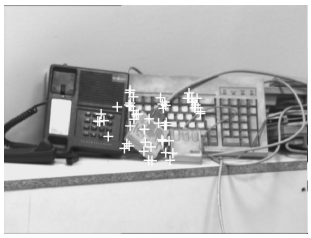
Image de référence

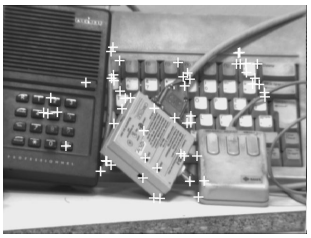
Image 8

FIG. 4: *Premier et dernier couple de la séquence Zoom.*

TAB. 3: Résultats de la séquence Zoom.

Image	Facteur d'échelle	Nombre d'appariements	Couples corrects	Pourcentage
1	0.42	27	27	100.0
2	0.48	88	88	100.0
3	0.62	101	101	100.0
4	0.79	185	185	100.0
5	1.19	149	149	100.0
6	1.47	173	171	98.8
7	1.85	111	110	99.1
8	2.1	59	55	93.2

7.4 Déformation perspective

Nous présentons enfin les résultats obtenus dans le cas d'une inclinaison du plan support. L'angle d'inclinaison entre l'image de référence et les images tests varie approximativement de -55^o à 55^o.

Pour la première et dernière image, le pourcentage de couples corrects est respectivement 74% et 89%. Si l'angle d'inclinaison est inférieur à 45^o le pourcentage de couples corrects est toujours supérieur à 95%.

8 Conclusion et perspectives

La méthode présentée dans cet article s'inscrit de le domaine très vaste de la mise en correspondance. Les primitives mises ici en correspondance sont des points extraits de zones fortement texturées. Ces points sont représentés

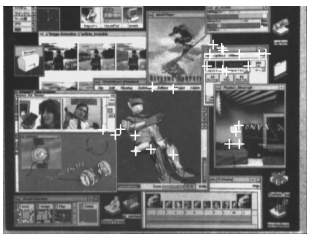
Image de référence


Première image


Image de référence

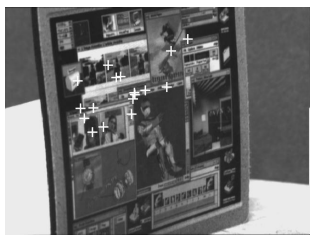
Dernière image

FIG. 5: *Premier et dernier couple de la séquence Perspective*

par un vecteur de caractéristiques locales directement basées sur le signal de luminance. Ce vecteur est invariant pour les rotations et translations de l'image et facilement intégrable dans un contexte multi-échelles.

Cette méthode de caractérisation est utilisé dans pour la reconnaissance et la localisation d'objets texturés. L'algorithme a déja été testé sur 4 objets texturés. Le nombre de primitives présentes dans la base de données est supérieur à 40000. Grâce à l'utilisation d'une technique d'indexation pour la mise en correspondance, le temps de recherche des vecteurs semblables est de l'ordre de la seconde pour 1000 points extraits de l'image à analyser.

Références

[1] R. Deriche. Using canny's criteria to derive a recursively implemented optimal edge detector. *International Journal of Computer Vision*, 1(2):167–187, 1987.

[2] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceeding of the 4th Alvey Vision Conference*, pages 147–151, 1988.

[3] J.J Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[4] H.P. Moravec. Visual mapping by a robot rover. In *Proceeding of the 6th International Joint Conference on Artificial Intelligence*, pages 598–600, 1979.

[5] C. Schmid. *Appariement d'images par invariants locaux de niveaux de gris.* Thèse de doctorat, Institut national polytechnique de Grenoble, 1995.