Mise en correspondance de descripteurs géométriques locaux par méthode *a contrario*

Julien Rabin, Julie Delon, Yann Gousseau

LTCI (UMR CNRS 5141), GET/TELECOM-PARIS
46 rue Barrault, 75013 Paris, France
Tél. 01 45.81.73.27 - fax. 01 45.81.71.44
{rabin,delon,gousseau}@tsi.enst.fr

Résumé – De nombreuses applications en analyse d'images s'appuient sur une représentation par des descripteurs locaux tels que les SIFT [3]. La mise en correspondance de ces descripteurs, bien que cruciale, est le plus souvent réduite à un seuillage sur la distance au plus proche voisin. Dans cette contribution, une nouvelle mesure de dissimilarité robuste à la quantification des descripteurs est proposée. Nous présentons ensuite un critère de mise en correspondance, inspiré des méthodes « *a contrario* » [1], qui permet d'évaluer le degré de significativité des appariements testés et fournit des seuils de validation qui s'adaptent automatiquement à la complexité et à la diversité des données.

Abstract – SIFT-like methods, based on the extraction and representation of local features, outperform global methods in object recognition and classification. These applications use a matching procedure between keypoint descriptors of several images, which -in practice- boils down to a closest-neighbor distance thresholding. A new dissimilarity measure is first introduced that improves descriptor quantization robustness. Then, an original matching criterion is presented, using "*a contrario*" statistical tests introduced in [1], which orders matches by meaningfulness and yields thresholds automatically adapted to the database.

1 Introduction

La représentation des images par des descripteurs locaux, en particulier par les descripteurs SIFT [3], s'est imposée dans nombre d'applications telles que la mise en correspondance d'images, la classification, ou encore la détection d'objets. Cette représentation est en effet plus robuste à certaines transformations et altérations de l'image que les approches globales. Elle nécessite d'abord l'extraction de points d'intérêt dans les images, puis la construction de descripteurs autour de ces points. Parmi tous les descripteurs proposés dans la littérature, une étude comparative [7] a montré la supériorité des descripteurs SIFT.

Ainsi, des images peuvent être comparées en mettant en correspondance les descripteurs qui les représentent. Plus formellement, cette comparaison consiste à apparier des descripteurs *requêtes* $\{a^i\}_{i=1...N_A}$ (extraits d'une image requête) et des descripteurs *candidats* $\{b^j\}_{j=1...N_B}$ d'une base de données (extraits d'une ou plusieurs images). Pour chaque requête a^i, les éléments b^j sont ordonnés selon leur similarité avec a^i. Un critère est alors utilisé pour décider quels candidats sont appariés avec a^i. Idéalement, ce critère doit maximiser le taux de bonnes détections, afin de faciliter et d'assurer la réussite d'éventuelles étapes postérieures (classification, groupement, etc).

Peu d'études portent sur cette étape de mise en correspondance. La méthode la plus utilisée en pratique, due à D. Lowe [3], consiste à comparer le rapport des distances entre a^i et ses deux plus proches voisins avec un seuil r. Ce critère donne souvent de bons résultats, mais souffre de plusieurs handicaps, que nous détaillerons dans la partie 4.

Pour pallier ces défauts, nous proposons d'utiliser une approche *a contrario* [2]. Le critère proposé (qui constitue la principale contribution de cet article) est décrit en détail dans la partie 4. Il fournit des seuils de décision qui s'adaptent automatiquement à la complexité de la requête et à la diversité de la base de données. Il permet également d'évaluer le degré de significativité des mises en correspondance et rend possible les détections multiples.

2 Extraction de caractéristiques

Cette partie présente brièvement la manière dont les points d'intérêt sont détectés et les descripteurs que nous construisons.

Points d'intérêt. La méthode « Laplace-Harris » que nous utilisons permet de ne garder que les structures en « coins », c'est à dire de courbure suffisamment grande. Une image I_0 est analysée *via* sa représentation $\{I_{\sigma_k}\}$ en espace multi-échelle linéaire, obtenue par convolution avec un noyau gaussien. Une recherche des maxima locaux de la réponse à l'opérateur laplacien normalisé [4] est ensuite effectuée en espace-échelle. Dans le but d'éliminer les structures de bord, on utilise le critère de Harris adapté à l'échelle de chaque point [6].

Orientations principales. Afin d'obtenir une représentation invariante par rotation, il est d'usage [3] d'attribuer à chaque point d'intérêt des orientations principales, extraites de l'histogramme circulaire d'orientation des gradients au voisinage du point. Pour cela, nous utilisons une méthode de sélection des modes significatifs d'un histogramme proposée par [2], que nous avons adaptée aux histogrammes circulaires. Le barycentre de chaque mode dé-

tecté définit une orientation principale du point d'intérêt. Cette extraction d'orientations est beaucoup plus robuste qu'une simple sélection des maxima de l'histogramme.

Construction du descripteur. Un descripteur est ensuite construit pour chacune de ces orientations. S'inspirant des SIFT, un descripteur s'écrit $a = (a_1, \ldots, a_M)$ où a_m est l'histogramme d'orientation du gradient (pondéré par sa norme) calculé sur l'un des M secteurs d'un masque autour du point d'intérêt (voir fig. 1). Chaque a_m a pour origine l'orientation principale du point considéré, est quantifié sur N cases et normalisé. Les gradients sont calculés à partir de l'image originale convoluée avec un noyau gaussien afin de se placer à l'échelle de la structure.

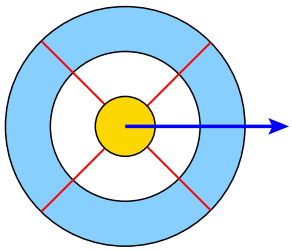


FIG. 1 – Masque de M (par défaut $M = 9$) secteurs utilisé pour la construction du descripteur. On utilise ici un masque circulaire afin d'être plus robuste aux rotations [7].

3 Mesure de dissimilarité

Pour comparer deux descripteurs, il est important d'utiliser une distance robuste à la quantification d'histogrammes et aux petits décalages angulaires. Les distances usuelles entre histogrammes (Euclidienne, Manhattan, Mahalanobis...) comparent des cases (*bins*) de mêmes indices et sont par conséquent très sensibles à la quantification. La métrique « Earth Mover Distance » EMD [9] (ou distance de transport) permet de s'affranchir de cette limitation en comparant des cases d'indices différents. Contrairement à ce qui est proposé dans [5] pour comparer des descripteurs SIFT, seuls les histogrammes de secteurs identiques sont ici comparés deux à deux, ce qui limite beaucoup la complexité algorithmique.

3.1 Distance de transport EMD entre histogrammes circulaires

Soient f et g deux histogrammes circulaires de N cases et normalisés : $\sum_{i=1}^{N} f[i] = \sum_{i=1}^{N} g[i] = 1$. Dans le cas non circulaire, on sait [11] que la distance EMD entre deux histogrammes est égale à la distance L^1 entre leurs histogrammes cumulés $\|F - G\|_1$. Dans le cas circulaire, on peut montrer que la distance EMD entre f et g est le minimum en k des distances L^1 entre les histogrammes F_k et G_k cumulés circulairement à partir de la k-ième cellule de quantification, *i.e.*

$$d(f,g) = \min_{k\in\{1,\ldots,N\}} \left\{ \frac{1}{N} \sum_{i=1}^{N} |\, F_k[i] - G_k[i]\,| \right\}, \text{ où} \quad (1)$$

$$\forall\, k,\ \ F_k[i] = \begin{cases} \displaystyle\sum_{j=k}^{i} f[j] & \text{si} \quad i \geq k \\ \displaystyle\sum_{j=k}^{N} f[j] + \sum_{j=1}^{i} f[j] & \text{si} \quad i < k \end{cases} . \quad (2)$$

Un descripteur, tel que nous l'avons défini dans la partie 2, est constitué de M histogrammes normalisés. On définit la mesure de dissimilarité entre deux descripteurs a et b par la somme des distances entre les histogrammes normalisés de même secteur m,

$$D(a,b) = \sum_{m=1}^{M} d(a_m, b_m)\,. \quad (3)$$

Cette mesure présente l'avantage d'être moins sensible au contexte que $\sum d(a_m,b_m)^2$ ou $\max d(a_m,b_m)$.

3.2 Performances

Pour quantifier l'intérêt de cette mesure de dissimilarité, les descripteurs d'une image sont mis en correspondance avec ceux d'une autre image, transformation affine de la première. Seul le plus proche voisin en dessous d'un seuil sur la distance est sélectionné. En faisant varier ce seuil, on obtient deux courbes (fig. 2) pour chaque distance (EMD et L^2) suivant deux quantifications (N=12 et 24). Elles montrent l'évolution du nombre de bonnes détections en fonction du nombre de fausses détections. On constate la supériorité de la distance EMD : la proportion de bonnes détections est meilleure et, contrairement à la distance L^2, elle augmente lorsque la quantification est plus fine.

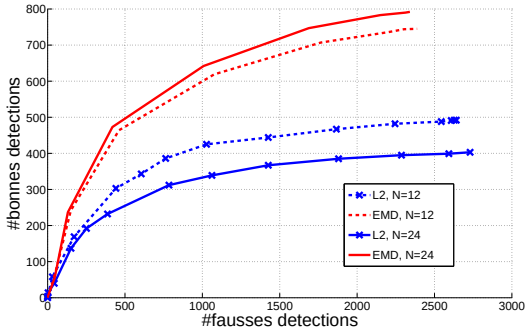


FIG. 2 – Comparaison des distances L^2 et EMD.

4 Mise en correspondance

L'étape de mise en correspondance consiste, pour chacun des descripteurs « requêtes » $a^i \in \{a^1, \ldots, a^{N_A}\}$, à décider avec quels descripteurs « candidats » $\{b^1, \ldots, b^{N_B}\}$ il y a mise en correspondance. Le critère le plus simple consiste à seuiller directement la mesure de dissimilarité. Il reste difficilement utilisable, essentiellement parce que les seuils conduisant à des résultats visuellement satisfaisants varient fortement d'une image à l'autre, mais également d'un descripteur requête à l'autre. Le raffinement le plus utilisé en pratique, dû à D. Lowe [3], consiste à calculer, pour chaque a^i, la distance à chacun des b^j. Si le rapport des distances au premier et au second plus proche voisin est inférieur à un seuil de détection r, **seul le plus**

proche voisin est apparié avec a^i. L'idée est que la restriction au plus proche voisin évite les appariements erronés multiples. Bien que ce critère (que l'on notera **NN2**) donne souvent de très bons résultats, il souffre des handicaps suivants :

– seuls les deux plus proches voisins sont retenus pour caractériser le contenu de la base de données. Il s'agit donc d'une forme d'apprentissage relativement pauvre de la complexité de la base et des spécificités de la requête.

– il n'y a pas de détection possible d'une structure présente plus d'une fois dans la base. Ceci est particulièrement limitant dans le cas de la détection d'objets multiples (par exemple lors de recherches dans une base de données volumineuse), mais également dans le cas, fréquent en pratique, d'objets présentant des structures répétitives (objets manufacturés, façades de bâtiments, etc.).

– le réglage optimal de r (fixé par l'utilisateur) est obtenu *a posteriori* car il varie d'une expérience à l'autre.

Afin de résoudre ces difficultés, nous utilisons l'approche ***a contrario*** introduite par [1].

4.1 Principe de la méthode *a contrario*

La méthode *a contrario* a été initialement proposée pour regrouper des caractéristiques visuelles de bas niveau. Elle a été utilisée avec succès dans les applications, entre autres, de détection d'alignements, de points de fuite et de bords contrastés [1, 2]. L'idée principale de cette méthode est de valider des groupes en rejetant une hypothèse d'indépendance des structures à grouper. Nous adaptons cette approche au problème de la mise en correspondance de descripteurs locaux, en introduisant un « modèle de fond » semblable à celui utilisé pour mettre en correspondance des morceaux de lignes de niveau dans [10].

4.2 Le modèle de fond

L'approche proposée consiste à valider les mises en correspondance pour lesquelles $D(a^i, b^j) = \sum_{m=1}^{M} d(a^i_m, b^j_m)$ est anormalement petite sous l'hypothèse $\mathcal{H}_0^i$ que les M distances $d(a^i_m, b^j_m)$ sont indépendantes. Pour i fixé, un descripteur aléatoire b est dit suivre un modèle de fond si l'hypothèse suivante est vérifiée :

$\mathcal{H}_0^i$: « les M variables $d(a^i_m, b_m)$ sont indépendantes ».

La probabilité que la dissimilarité entre a^i et un élément b de la base soit bornée par un seuil δ sous l'hypothèse $\mathcal{H}_0^i$ s'écrit alors

$$\mathbb{P}\left(D(a^i, b) \leq \delta \,|\, \mathcal{H}_0^i\right) = \int_{-\infty}^{\delta} \underset{m=1}{\overset{9}{*}} p^i_m(x)\, dx \,, \qquad (4)$$

où $*$ est le produit de convolution et p^i_m la densité de probabilité de $d(a^i_m, b_m)$. Afin d'apprendre le modèle de fond sur la base de données et de prendre en compte la spécificité de chaque a^i, chaque marginale p^i_m est obtenue directement à l'aide de la distribution empirique de $d(a^i_m, b_m)$ calculée sur $\{b^1_m, \ldots, b^{N_B}_m\}$.

4.3 Nombre de fausses alarmes

L'étape suivante consiste à fixer un seuil de détection sur la probabilité $\mathbb{P}(D(a^i, b) \leq \delta \,|\, \mathcal{H}_0^i)$. Le réglage d'un tel seuil est délicat, car il nécessite un compromis entre le taux de bonnes et de mauvaises détections. Afin de fixer ce seuil, nous introduisons la fonction « NFA » et le seuil $\tilde{\delta}_i(\varepsilon)$:

$$\mathrm{NFA}(a^i, \delta) = N_A\, N_B\, \mathbb{P}(D(a^i, b) \leq \delta \,|\, \mathcal{H}_0^i) \,, \qquad (5)$$

$$\tilde{\delta}_i(\varepsilon) = \arg\max_{\delta} \left\{ \mathrm{NFA}(a^i, \delta) \leq \varepsilon \right\} \,. \qquad (6)$$

La variable $\tilde{\delta}_i(\varepsilon)$ représente le seuil sur la mesure de dissimilarité du descripteur a^i qui permet de ne valider que ses correspondances avec la base qui ont un NFA borné par ε. On montre facilement qu'avec un tel critère, lorsque l'on teste les $N_A \times N_B$ correspondances possibles, l'**espérance** du nombre de mise en correspondance sous l'ensemble des hypothèses $\mathcal{H}_0^i$ (appelé *nombre de fausses alarmes*) est borné par ε. Le seuil $\tilde{\delta}_i(\varepsilon)$ est **automatiquement** déterminé à partir de ε et varie d'un descripteur a^i à un autre. La quantité $\mathrm{NFA}(a^i, D(a^i, b^j))$ permet par ailleurs *in fine* d'évaluer la significativité des correspondances des a^i avec les b^j et de les trier suivant ce critère.

Interprétation en termes de tests statistiques. Pour chaque i, j on définit le test permettant de rejeter l'hypothèse nulle $\mathcal{H}_0^i$ lorsque $D(a^i, b^j) \leq \tilde{\delta}_i(\varepsilon)$. On retient la mise en correspondance de a^i et b^j lorsque $\mathcal{H}_0^i$ est rejetée. Le seuil $\tilde{\delta}_i(\varepsilon)$ est fixé de manière à contrôler le risque de première espèce. Le terme $N_A N_B$ apparaissant dans (5) correspond à une correction dite de Bonferonni dans le cadre des tests multiples [8].

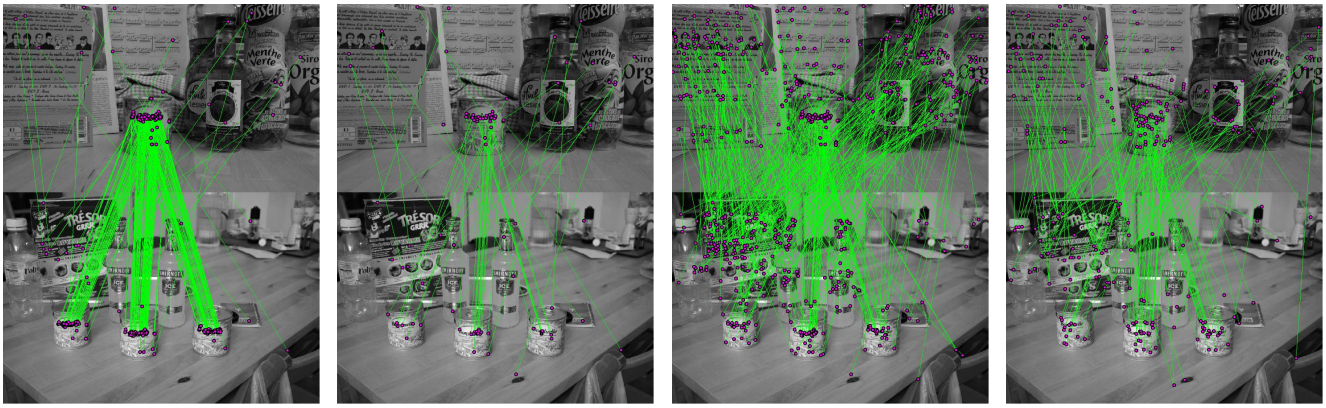
5 Résultats

À partir de ε imposé par l'utilisateur, les mises en correspondance sont sélectionnées et triées selon leur pertinence, les seuils $\tilde{\delta}_i(\varepsilon)$ étant évalués indépendamment pour chacun des N_A points d'intérêt de l'image, ce que ne permet pas un seuil δ fixé *a priori* sur la distance. Le choix du seuil ε correspond au degré de fiabilité des mises en correspondance. En pratique, on choisit le seuil à $\varepsilon = 10^{-1}$ afin de limiter le nombre de correspondances perçues comme des fausses détections (des objets appariés différents), bien que ce ne soit pas toujours des fausses alarmes (les structures sont bien similaires).

Nous comparons ici les résultats de trois méthodes de mise en correspondance : l'approche originale de D. Lowe [1] [3] notée SIFT-NN2 (distance euclidienne sur les SIFT puis critère NN2) et les approches EMD-NN2 et EMD-NFA qui utilisent les descripteurs décrits dans la partie 2, la distance EMD et respectivement le critère NN2 et notre critère *a contrario* (simplement noté NFA).

La première expérience (fig. 3) consiste à retrouver un objet (une boîte de conserve) présent trois fois dans une scène. Les figures 3(a), 3(b) et 3(c) montrent les appariements obtenus respectivement par EMD-NFA à $\varepsilon = 10^{-1}$ et par EMD-NN2 à $r = 0.8$ puis à $r = 0.9$. On peut constater que notre méthode EMD-NFA permet d'ajuster automatiquement le seuil sur la distance de manière à ne retenir

[1] dont l'algorithme est gracieusement mis à disposition depuis `http://www.cs.ubc.ca/~lowe/keypoints/`

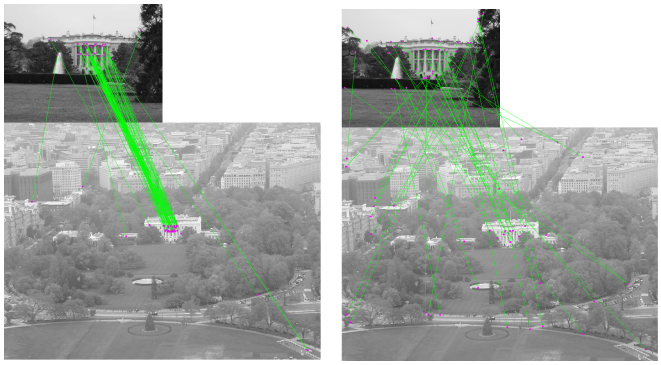


(a) EMD-NFA à $\varepsilon = 10^{-1}$ (b) EMD-NN2 à $r = 0.8$ (c) EMD-NN2 à $r = 0.9$ (d) SIFT-NN2 à $r = 0.8$

FIG. 3 – Le critère NFA ajuste automatiquement les seuils de détection pour mettre en correspondance les points de l'objet simultanément avec les 3 boîtes présentes dans la scène (208 bonnes détections parmi 228). Il permet également de limiter le taux de fausses détections contrairement au critère NN2 (10% de fausses détections avec EMD-NFA à $\varepsilon = 10^{-1}$, 36% avec EMD-NN2 à $r = 0.8$, 80% avec EMD-NN2 à $r = 0.9$ et 59% avec SIFT-NN2 à $r = 0.8$).

que les mises en correspondance avec les trois objets, avec seulement quelques fausses détections (20 sur un total de 228). Avec $r = 0.8$, le critère NN2 fournit un nombre de fausses détections identique mais pour seulement 55 appariements. En augmentant le seuil de détection à $r = 0.9$, on n'obtient que 42 bonnes détections supplémentaires au prix d'un accroissement important du nombre de fausses détections (301 parmi un total de 378). Il en va de même avec SIFT-NN2 (avec le seuil $r = 0.8$ préconisé dans [3]) sur la figure 3(d), où 97 mises en correspondances sont erronées sur un total de 165.

La deuxième expérience (fig. 4) illustre la difficulté de mettre en correspondance des structures répétitives telles que les colonnes de la façade de la Maison Blanche avec le critère NN2. Les figures 4(a) et 4(b) montrent respectivement les correspondances obtenues avec EMD-NFA à $\varepsilon = 10^{-1}$ et SIFT-NN2 à $r = 0.8$.



(a) EMD-NFA à $\varepsilon = 10^{-1}$: 68 bonnes détections sur 73 (b) SIFT-NN2 à $r = 0.8$: 8 bonnes détections sur 41

FIG. 4 – Le critère *a contrario* (fig. 4(a)) permet d'obtenir beaucoup plus de bonnes détections sur la façade de la maison blanche que le critère NN2 (fig. 4(b)).

6 Conclusion et perspectives

Cette contribution présente une procédure de mise en correspondance de descripteurs locaux originale, qui repose d'une part sur une distance robuste entre descripteurs et d'autre part sur un critère de décision inspiré des méthodes *a contrario*. Ce critère fournit des seuils de décision automatiques, adaptés à la diversité des données, et ne borne pas *a priori* le nombre d'appariements pour chaque requête. Il présente également l'avantage d'être générique : nous travaillons actuellement à son adaptation dans le cas de descripteurs utilisant la couleur. La poursuite naturelle de ce travail sera la prise en compte de la cohérence globale des appariements, méthode classique en reconnaissance d'objets. Dans ce but, la même méthodologie *a contrario* pourra être utilisée, afin de s'adapter à la taille et à la complexité des bases de données.

Références

[1] A. Desolneux, L. Moisan, et J.-M. Morel. Meaningful Alignments. IJCV, 40(1), p. 7-23, 2000.

[2] A. Desolneux, L. Moisan, et J.-M. Morel. Gestalt Theory and Image Analysis. Lecture Notes, Springer, 2007.

[3] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. IJCV, p. 91-110, 2004.

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

[5] H. Ling and K. Okada. EMD-L1 : An Efficient and Robust Algorithm for Comparing Histogram-Based Descriptors. ECCV, p. 330-343, 2006.

[6] K. Mikolajczyk et C. Schmid. Indexing Based on Scale Invariant Interest Points. ICCV, 2001

[7] K. Mikolajczyk et C. Schmid. A Performance Evaluation of Local Descriptors. CVPR, 2003.

[8] R. G. Miller. *Simultaneous Statistical Inference*. Springer-Verlag, New York, 1991.

[9] Y. Rubner, C. Tomasi, et L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. IJCV, p. 99-121, 2000.

[10] F. Sur, P. Musé, Y. Gousseau, F. Cao et J-M. Morel. An a contrario Decision Method for Shape Element Recognition. IJCV, 69(3), p. 295-316, 2006.

[11] C. Villani. *Topics in Optimal Transportation*. AMS, 2003.