Les représentations « steerables » pour la compression et l'extraction de caracteristiques de bas niveau.

François Tonnin, Patrick Gros, Christine Guillemot

Irisa, Avenue du Général Leclerc, 35042 Rennes, France
ftonnin@irisa.fr

Résumé – L'extraction de caractéristiques de bas niveau s'effectue à partir de la représentation des images dans l'espace-échelle Gaussien. C'est l'unique représentation satisfaisant à certaines conditions permettant d'obtenir des descripteurs invariants aux similitudes du plan. Toutefois, sa forte redondance et sa nature non creuse empêchent toute compression. Le calcul de nouveaux descripteurs nécessite donc de décompresser la base d'images puis de la transformer dans l'espace-échelle Gaussien. Dans ce papier, nous proposons un descripteur local dans le domaine compressé donné par les coefficients quantifiés des transformées « steerables ». La robustesse et le pouvoir discriminant sont comparés à ceux existants dans l'espace-échelle Gaussien et évalués en fonction de l'entropie de la représentation.

Abstract – Low-level features are extracted in the Gaussian scale-space of images. It is the unique representation satisfying some conditions aiming at extracting features invariant to translations, rotations and scale changes. Yet, it is higly redundant and non sparse so that compression schemes can not be designed in this representation. In this paper, we propose to extract local features directly in the compressed domain provided by quantized coefficients of steerable transforms. The robustness and discriminative power of the features are compared to state-of-the-art features, and given in function of the entropy of the representation.

1 Introduction

La transformée de Fourier et la transformée de Gabor constituaient jusqu'à une époque récente les seules alternatives à la représentation en niveaux de gris des images. Il existe aujourd'hui de nombreuses autres représentations, comme la transformée en pyramide Laplacienne, les transformées en ondelettes, en curvelets, en bandelets. Ces nouvelles représentations ont permis des progrès considérables en compression et débruitage. Étant à échantillonnage critique ou de faible redondance, elles sont en revanche inadaptées à l'extraction robuste de caractéristiques de bas niveau comme des points, des contours, ou des descripteurs visuels locaux. Une telle extraction constitue un pré-traîtement pour de nombreuses tâches visuelles, comme la segmentation, la détection de mouvement, le suivi ou la reconnaissance d'objet, ou la recherche d'images par le contenu. L'espace-échelle Gaussien est l'unique représentation utilisée pour effectuer ces traîtements [2, 3, 6, 8, 7]. La gestion de grandes bases d'images est fortement pénalisée, en termes de flexibilité et de temps de calcul, par la nécessité de d'abord décompresser les images, puis de les convertir dans l'espace-échelle Gaussien, avant d'extraire les caractéristiques de bas niveau requises. Un problème d'intérêt consiste donc à trouver des représentations d'images adaptées à la compression et à la description locale des images. Toutefois, compression et description peuvent paraître antagonistes.

Dans un but de description, la représentation doît être robuste à un ensemble de transformations admissibles, idéalement constitué de l'ensemble des changements de perspective.Réduisant cet ensemble au groupe des similitudes du plan (groupe des translations, rotations et homothéties), et ajoutant la contrainte de causalité, la représentation est déterminée. Il s'agit de l'espace-échelle Gaussien [11].

En revanche, dans un souci de compression, l'image doît pouvoir être reconstruite à partir d'une quantité minimale d'information fournie par les coefficients quantifiés de sa transformée. La représentation doît donc être creuse, à échantillonnage critique (ou de redondance minimale), et idéalement à coefficients indépendants.Une représentation creuse est obtenue en adaptant les fonctions de projection sur l'espace transformé aux statistiques des images naturelles. L'indépendance entre coefficients est impossible, aussi préfère-t-on des fonctions de projection orthogonales, de manière à obtenir des coefficients décorrélés. Une représentation multirésolution à échantillonnage critique n'est pas souhaitable. En effet ce type de représentation est variant aux translations, donc inadapté à l'extraction robuste de caractéristiques de bas niveau. Certaines représentations multirésolution [4, 10] ont été spécialement conçues pour préserver la covariance aux translations et rotations. Le prix à payer est une certaine redondance dans la représentation. Dans

certains cas, il existe des techniques [1, 12] pour réduire le coût de codage dû à cette redondance.
Dans ce papier, nous proposons un descripteur local dans le domaine compressé donné par les coefficients quantifiés des transformées « steerables ».
La section 2 définit le protocole expérimental permettant d'évaluer la robustesse des points extraits, et la discrimination des descripteurs visuels. L'extracteur de points est décrit dansla section 3, et les descripteurs locaux dans la section 4. Un test de recherche d'images dans une base de 16000 images est discuté en conclusion, ainsi que la qualité des caractéristiques extraites en fonction de l'entropie de la représentation.

2 Protocole expérimental

L'extraction robuste de points est un traîtement requis par de nombreuses tâches visuelles.Ces points sont robustes à un ensemble $\mathcal{T}$ de transformations admissibles, usuellement réduits à l'ensemble des similitudes du plan. De tels points sont appelés points d'intérêt.
Se donnant une image naturelle $I_0 \in l^2(\mathbb{Z}^2)$ (l'ensemble des suites à valeur dans $\mathbb{Z}^2$ et de carré sommable), les points d'intérêt sont extraits à partir de la représentation R_0 de l'image I_0. Cette représentation R_0 devant être covariante aux translations et changements d'échelle, elle est nécessairement donnée par la convolution entre l'image I_0 et un filtre paramétré en échelle:

$$R_0(x, y; s) = I_0(x, y) \star h_s(x, y) \ , \qquad (1)$$

où l'échelle s est à valeur dans $\mathbb{R}_+$. Les représentations multirésolution avec decimation ne sont pas compatibles avec cette définition. Elles seront néanmoins considérées, et dans le cadre de leur évaluation, les coordonnéés des points extraits seront *ramenées* sur la grille de l'image originale. L'ensemble P^0 des n_0 points d'intérêt extraits à partir de R_0 est noté

$$P^0 = \{(x_i^0, y_i^0; s_i^0), 1 \leq i \leq n_0\} \ , \qquad (2)$$

où (x_i^0, y_i^0) sont les coordonnées spatiales du $i^{ème}$ point extrait et s_i^0 son échelle. En vue d'évaluer la robustesse de ces points à une transformation admissible $T \in \mathcal{T}$, l'image synthetique $I_1 = T \circ I_0$ et sa representation R_1 sont créées. L'ensemble P^1 des n_1 points d'intérêt extraits à partir de la représentation n'est pas défini comme P_0, mais de la manière suivante:

$$P^1 = \{\left(T^{-1}(x_i^1, y_i^1); s_i^{1\to 0}\right), 1 \leq i \leq n_1\} \ , \qquad (3)$$

où l'échelle $s_i^{1\to 0}$ est égal à l'échelle s_i^1 divisée par le facteur d'échelle de la transformation T. La robustesse des points d'intérêt à une transformation admissible $T \in \mathcal{T}$ est évaluée à travers la notion de répétabilité, introduite dans [9], et définie par la proportion de points qui se correspondent:

$$r = \frac{\max(|C^{01}|, |C^{10}|)}{\min(n_0, n_1)} \ , \qquad (4)$$

où $|C^{01}|$ est le cardinal de C^{01}, et C^{01} (resp. C^{10}) est le sous ensemble des points de P^1 (resp. de P^0) qui ont un correspondant dans P^0 (resp. dans P^1). Dans [9], ce sous ensemble est défini relativement à une précision ε:

$$C^{01} = \{(x^1, y^1; s^1) \in P^1 : \\ \exists (x^0, y^0; s^0) \in P^0 \mid d\left((x^0, y^0), (x^1, y^1)\right) \leq \varepsilon\} \ , \qquad (5)$$

où d est la distance euclidienne. Ainsi définie, la répétabilité a tendance à croître avec le nombre de points extraits, et à décroître avec la taille de l'image. Plus précisément, l'espérance de répétabilité de n points aléatoirement extraits à partir d'une image composée de N pixels est $r(\varepsilon) = \frac{n\varepsilon^2}{N}$. Dans la suite, le nombre de points extraits est fixé à 400, et la taille des images est 480×320 pixels.
Les points d'intérêt peuvent servir à calculer des descripteurs visuels locaux. Un descripteur local est calculé à partir du voisinage d'un point extrait. Idéalement,il caractérise de façon unique l'ensemble des voisinages déformés par les transformations admissibles $T \in \mathcal{T}$. En pratique, se donnant un point d'intérêt $(x_i^k, y_i^k; s_i^k)$, on cherche un descripteur discriminant m_i^k, $(k \in \{0, 1\})$ invariant aux transformations admissibles. Habituellement,l'invariance et le pouvoir discriminant sont simultanément évalués en mesurant la performance d'un système de recherche d'images par le contenu. Dans la suite, ces deux caractéristiques seront évaluées par la répétabilité obtenue en appariant les points dont les descripteurs sont les plus proches au sens de la distance euclidienne. Le sous ensemble C^{01} des points de P^1 appariés de cette manière à la precision ε, est définie par

$$C^{01} = \{(x^1, y^1; s^1) \in P^1 : \mathrm{d}\left((x_j^0, y_j^0), (x^1, y^1)\right) \leq \varepsilon, \\ j = \arg \min_{1 \leq p \leq n_0} ||m^1 - m_p^0||\}. \qquad (6)$$

Cette répétabilité est majorée par la répétabilité définie par Eqn. 5. Le ratio entre ces deux répétabilités est dans la suite appelé *répétabilité par descripteurs.*

3 Détection de points

Cette transformée, présentée dans [10], satisfait à plusieurs propriétés la rendant intéressante pour le problème conjoint de compression et de description. Elle est quasi invariante aux translations, permet une fine analyse angulaire, et enfin il existe des techniques permettant de réduire le coût de codage dû à sa redondance.
Elle est conçue dans le domaine de Fourier, où le noyau h_s défini dans Eqn.1 est séparable, i.e. sa transformée de Fourier est de la forme $\hat{h}_s(\rho, \theta) = U_s(\rho)V(\theta)$. Le noyau $U_s(\rho)$ est passe bande et assure que l'énergie globale de chaque bande est invariante aux translations. Le noyau $V(\theta)$ est « steerable », i.e. il existe un ensemble d'angles d'analyse

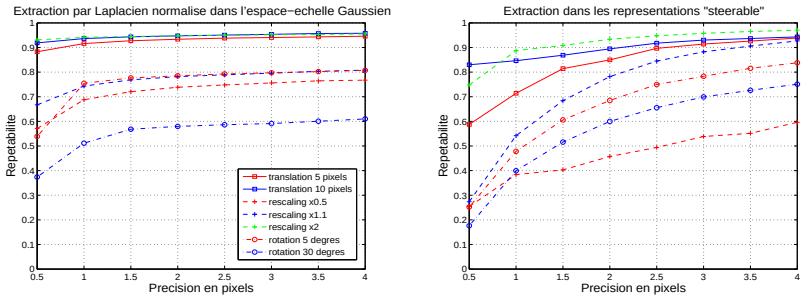


FIG. 1: *Répétabilité des points extraits en espace-échelle Gaussien par l'extracteur de Lowe (à gauche), et dans les représentations "steerables"(à droite).*

$\{\alpha_k\}_k$, et un ensemble de fonctions d'interpolation $\{f_k\}_k$ tels que:

$$\forall\theta \in [0,\pi],\ V(\theta) = \sum_{k=0}^{N-1} f_k(\theta) V(\alpha_k) \qquad (7)$$

Cela signifie qu'à une position et une échelle fixées, les N coefficients calculés pour chacun des angles d'analyse sont suffisants pour reconstruire la transformée en tout angle. A travers les échelles, la représentation est pyramidale avec un facteur quatre de décimation entre deux échelles, la première octave n'étant pas décimée. Choisissant $N = 4$ angles de base, la redondance de cette représentation est de l'ordre de $\frac{16}{3}$. Les points extraits sont les plus forts maxima locaux en position et orientation, mais non pas en échelle, à cause de la trop grossière discrétisation en échelle. Fig. 1 montre que les points extraits sont plus répétables mais moins précis que dans [7]. Une localisation sous-pixellique et l'élimination des points de courbure trop faible permettraient d'améliorer la précision. Les représentations « steerables » offrent donc un bon compromis entre redondance et répétabilité. De plus, la fine analyse angulaire effectuée par ce type de transformée permet la conception des descripteurs visuels locaux présentés dans la prochaine section.

4 Description locale

Considérant que la représentation d'images utilisée est covariante à toute similitude du plan, et que les points extraits sont parfaitement répétables, le voisinage des points extraîts n'est affecté que par la rotation et dans une moindre mesure par le changement d'échelle. En effet, dans ce cas idéal, l'impact de changements d'échelle de facteur proportionnel au pas de discrétisation en échelle de la grille θ est même nul. Le problème principal lors de la conception de descripteurs locaux est donc l'invariance aux rotations. Une première solution est représentée par les invariants différentiels [5] qui sont des quantités invariantes aux rotations. Une seconde solution consiste à affecter à tout point extrait un angle robuste, puis de calculer le descripteur relativement à cette orientation principale. Le descripteur SIFT [7] en est un exemple. Il se transpose aisément dans les représentations « steerables ». L'extraction de l'orientation principale s'effectue à partir des pixels $\{x_i, y_j\}_{i,j}$ du voisinage 16×16 du point extrait (x, y). En chacun de ces points est calculée l'orientation maximisant l'énergie définie par Eqn. 1 et Eqn. 7:

$$\theta_{i,j} = \arg \max_{0\le p\le 35} R(x_i, y_j, s; \frac{p\pi}{18}) \ , \qquad (8)$$

et son poids associé est donné par

$$m_{i,j} = R(x_i, y_j; s; \theta_{i,j}) \qquad (9)$$

L'orientation principale affectée au point d'intérêt est celle où l'histogramme des orientations locales pondérées est maximale. Cette orientation est utilisée pour partitionner le voisinage 16×16 en 4×4 carré de taille 4×4 pixels. Le descripteur SIFT finalement calculé est constitué de la concaténation de chacun des histogrammes des orientations pondérées calculées dans chacun de ces carrés. Le graphique de gauche de Fig. 2 montre que l'appariement de points par leur descripteur SIFT est presque parfait. Le graphique du milieu montre une nette dégradation lorsque cet appariement s'effectue à partir des moment différentiels, et celui de droite donne un résultat intermédiaire pour notre transposition de SIFT aux représentations « steerables ». La dégradation s'explique par la redondance beaucoup plus faible de ces représentations par rapport à celle utilisée dans la version originale de SIFT. En particulier, l'image n'est pas interpolée sur une grille quatre fois plus fine lors du traîtement de la première octave. Comme le montrent les résultats présentés en conclusion, un compromis est à trouver entre la redondance de la représentation et la robustesse des caractéristiques extraites.

5 Conclusion

L'espace-échelle Gaussien est l'unique représentation utilisée pour l'extraction de caractéristiques locales de bas niveau. Son unicité provient des contraintes de covariance aux similitudes et de causalité. Si l'on relâche cette dernière contrainte, de nombreuses représentations sont possibles, parmi lesquelles les représentations « steerables ». Elles ont l'avantage d'être creuses et de faible redondance,

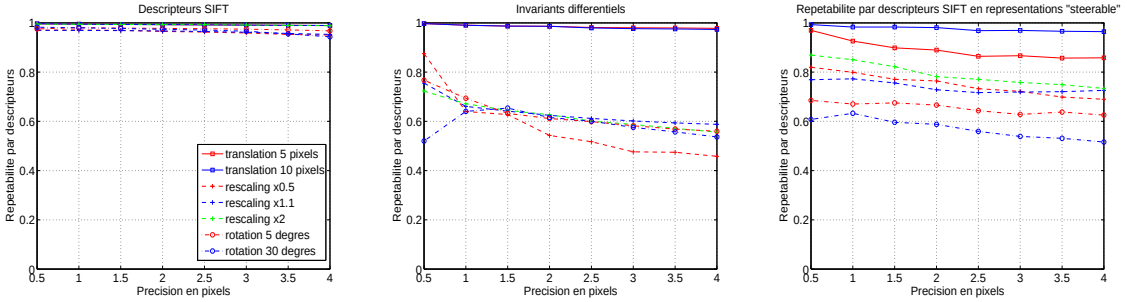
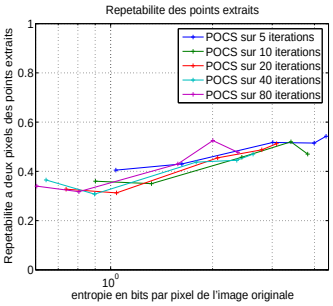
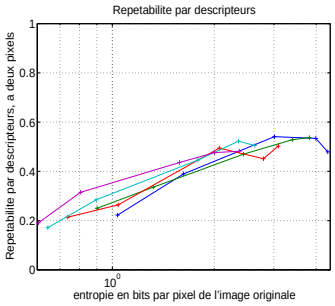


FIG. 2: *Répétabilité par descripteurs des descripteurs locaux SIFT, des invarariants différentiels, et du descripteur SIFT transposé dans les représentations "steerables".*





Tab.1	requête 1	requête 2
rot 5	735/15	1011/4
rot 30	660/9	825/4
scale x0.5	461/8	157/33

Tab.2	requête 1	requête 2
rot 5	282/23	354/10
rot 30	247/17	309/9
scale x0.5	186/19	98/46

FIG. 3: *Figures de répétabilité des points et par descripteurs en fonction de l'entropie de la représentation, et tableau des votes de recherche de copie (cf. conclusion)*

permettant ainsi de concevoir des schémas de compression. Compressant par POCS [12] avec différents seuils, et quantifiant uniformément sur 5 bits, Fig. 3 donne les répétabilités à deux pixels en fonction de l'entropie de la représentation, pour une rotation de 10 degrés. La dégradation est importante, mais pas sufisamment pour empêcher l'extraction des caractéristiques locales directement dans le domaine compressé. La capacité de retrouver l'image originale à partir de l'image transformée reste très bonne. Ce test est décrit par le tableau de Fig. 3. Tab.1 donne, pour une base de 16000 images de taille 480×320, le nombre de votes de la "bonne" image et le nombre de votes le plus élevé parmi les "mauvaises" images. Tab.2 donne la même information pour une base de 3000 images compressées et quantifiée à 2 bits par pixels de l'image originale.

Références

[1] B. Beferull-Lozano and A. Ortega. Coding techniques for oversampled steerable transforms. In *Proc. of thirty-third Intl. Asilomar Conf. on Signals, Systems and Computers*, 1999.

[2] J. Canny. A computational approach to edge detection. *Trans. on Pattern Analysis and Machine Intelligence*, 8(6):679–698, November 1986.

[3] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of Fourth Alvey Vision Conf.*, pages 147–151, 1988.

[4] N. Kingsburry. The dual-tee complex wavelet transform: a new efficient tool for image restoration and enhancement. In *European Signal Processing Conf.*, pages 319–322, 1998.

[5] J.J. Koenderinck and A.J. Van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55(6):367–375, March 1987.

[6] T. Lindeberg. *Scale-space theory in computer vision*. Kluwer Academic Publisher, 1994.

[7] D.G. Lowe. Object recognition from local scale-invariant features. In *Intl. Conf. on Computer Vision*, pages 1150–1157, 1999.

[8] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(5):530–534, May 1997.

[9] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *Intl. Conf. on Computer Vision*, pages 230–235, 1998.

[10] E.P. Simoncelli, W.T. Freeman E.H. Adelson, and D.J. Heeger. Shiftable multiscale transforms. *IEEE Trans. on Information Theory*, 38(2):587–607, March 1992.

[11] A.P. Witkin. Scale space filtering. In *Intl. Joint Conf. on Artificial Intelligence*, pages 1019–1022, 1983.

[12] D.C. Youla. Generalized image restoration by the method of alternating orthogonal projections. *IEEE Trans. Circuits and Systems*, 25(9):694–702, September 1978.