Extraction de points robustes dans des images naturelles complexes

François-Xavier Espiau
INRIA Sophia-Antipolis, projet ICARE
2004 route des Lucioles, BP 93
06902 Sophia-Antipolis
Francois-Xavier.Espiau@sophia.inria.fr

Résumé - *Si de nombreux progrès ont été faits dans le monde de la vision par ordinateur depuis quelques années, avec notamment l'apport de la géométrie projective, tous les problèmes n'en n'ont pas été résolus pour autant. En effet, si les scènes structurées sont de mieux en mieux traitées, les images naturelles, souvent non structurées, sont encore source de nombreux problèmes. Dans cet article, nous présentons une méthodologie pour robustifier les techniques d'extraction d'informations robustes dans le cas d'images naturelles complexes appliquée à l'imagerie sous-marine.*

Mots clé - images complexes, extraction de points d'intérêts

1 Introduction

Dans le traitement d'images naturelles, on essaie souvent de trouver des informations géométriques que l'on sait relativement bien traiter commes des coins ou des droites. Mais dans notre problématique, nous ne disposons que de très peu d'informations sur les scènes observées issues de séquences vidéos: pas de vision stéréo, pas de calibration de la caméra, pas de connaissances sur le mouvement de celle-ci, les objets ne sont pas structurés mais néanmoins rigides et leur texture est aléatoire. Posé ainsi, le problème d'extraction d'informations robustes dans un scène quelconque peut vite se réveler difficile à résoudre. Mais en le posant différemment, il est possible de traiter des images aussi complexes que celle présentée en figure 1. Dans un premier temps, nous allons voir ce que l'on peut et ce qu'il "faut" attendre comme type d'informations de cette catégorie d'images, puis nous détaillerons comment y parvenir grâce notamment à une approche multi-échelles et une bonne exploitation du détecteur de Harris. Nous expliquerons alors la notion d'appariement pyramidal qui a pour but de classifier les points d'intérêt détectés et de trouver les points les plus stables et robustes. Dans notre approche, nous intégrons également la notion de rigidité globale de la scène en utilisant un maillage de Delaunay que nous détaillerons. Enfin, nous présenterons les résultats obtenus sur des images réelles puis nous aborderons succinctement la suite des travaux.

FIG. 1 – *Fumerolle*

2 Etat de l'art

Le traitement d'images naturelles (souvent complexes à cause des textures, des formes ou encore des différents bruits) nécessite des algorithmes dits "bas-niveau" de ces images, ce qui amène généralement à extraire certaines informations comme des contours ou des coins. Ces techniques sont souvent basées sur le détecteur de Canny-Deriche [2], [3] pour les contours, et donc sur le gradient de l'image. Des techniques plus poussées comme les contours déformables [6] ont été menées plus récemment et se basent, elles, sur des modèles de forces physiques. Mais un des problèmes rencontrés comme celui du chaînage des contours, peut vite devenir délicat pour des images relativement peu structurées et bruitées. Pour les détecteurs de points, de nombreuses approches ont été proposées [1] [7] [4] [5] [12] [11] [16] et celles-ci se basent sur différentes représentations de coins théoriques, soit sur des dérivées secondes, soit des modèles de masques de pixels, soit sur différentes mesures, etc.

Certaines approches plus récentes basées sur la résolution d'équations aux dérivées partielles permettent de faire de la restauration d'images bruitées et/ou endommagées ou bien encore faire du suivi d'objets

[8]. Des approches reposant sur des techniques markoviennes sont également utilisées pour faire de la segmentation d'images satellitaires, comme par exemple segmenter des zones urbaines et péri-urbaines [10]. Mais ces méthodes s'appuient aussi sur de l'apprentissage (sur les textures dans le cas d'images satellites) et par conséquent ne sont pas utilisables dans notre cas.

Il existe aussi des approches basées sur la détection de mouvement dans des séquences d'images. Celles-ci sont souvent liées au flot optique [17] et du suivi dans des séquences vidéos: dans ces approches, on privilégie l'aspect temporel. A partir des différents mouvements contenus dans la séquence, on peut faire de la segmentation et retrouver différents objets. Mais ces techniques ne sont robustes que sous certaines contraintes fortes comme la rigidité des objets, des séquences temporelles relativement denses, ...

A partir des informations extraites, il est alors possible de faire de l'appariement entre images afin de faire de la reconstruction 3d. Les techniques sont nombreuses et un grand nombre utilisent différentes mesures de corrélation. Citons néanmoins deux approches intéressantes: celle de C. Schmid qui fait de l'appariement grâce à des vecteurs d'invariants en niveaux de gris [13] et celle de M. Lhuillier [9] qui construit des maillages de Delaunay dans l'image et utilise des triangles pour l'appariement.

3 Approche proposée

Après avoir mené une étude sur différents détecteurs de coins (SUSAN [11], CSS [16], Harris [5] modifié avec l'implémentation récursive des gaussiennes de Deriche [4] et la stabilisation du détecteur de Schmid [14]), nous avons porté notre choix sur ce dernier car il s'est révélé être plus précis en localisation, plus robuste au bruit et avec une répétabilité dans la détection plus forte que les autres.

L'image étant très bruitée (bruits de mesure liés aux capteurs, perturbations annexes) et ne connaissant pas de modèle de cet ensemble de bruits, nous devons néanmoins extraire des informations robustes.

L'approche par contours étant vouée à l'échec, nous devons extraire des points d'intérêts qui soient à la fois robustes au bruit et bien localisés. Mais cette information n'est que locale et même si les objets ne sont pas structurés, nous devons prendre en compte le fait qu'ils sont rigides: cela va se faire par le biais d'une triangulation de Delaunay. A noter que nous avons pris le parti de ne pas utiliser l'information temporelle dans cette approche.

3.1 Approche pyramidale

Pour chaque image issue d'une séquence vidéo, on construit une pyramide d'image. Celle-ci nous permet d'obtenir un ensemble de vues synthétiques de l'image originale. Le principe de construction est le suivant. Soit I_k la *k-ème* image de la séquence, on définit alors le niveau minimal 0 de la pyramide associée comme étant l'image I_k de résolution maximale et on la note I_k^0. Pour construire les images de la pyramide, on procède comme suit: $I_{k,l+1} = f(I_{k,l}) = h \circ (g * I_{k,l})$ où g est une fonction de lissage (ici une gaussienne), h fonction d'échantillonnage et $\circ$ et $*$ étant respectivement les opérateurs de composition et de convolution, qui s'écrit dans le cas discret et pour deux variables:

$$f(x,y) * g(x,y) = \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} f(k,l).g(x-k,y-l)$$

L'intérêt d'utiliser une gaussienne réside d'une part, dans le fait qu'en pratique, elle peut être très bien approximée par une fonction à support compact et donc réduire considérablement les calculs et d'autre part, que cette fonction élimine relativement bien le bruit.

Une conséquence de cette approche est que le lissage est isotrope et n'est donc pas adaptatif en fonction des textures. Pour des raisons de temps de calcul, nous n'avons pas retenu cette idée. Enfin, en pratique, on estime que l'imagette la plus petite exploitable sera de taille 64x64, le signal étant trop dégradé ensuite.

L'échantillonnage que nous appliquons est arbitraire (un pixel sur deux) et il existe donc plusieurs possibilités pour échantillonner une image. Cela peut être un phénomène générant des erreurs mais en pratique, il n'en est rien car, composé avec une fonction de lissage, le signal obtenu est quasi-identique et cela n'a pas de répercution notable sur les différents calculs effectués par la suite.

A l'issue de la construction de cette pyramide, nous avons donc obtenu de nouvelles images synthétiques qui conservent néanmoins les caractéristiques principales de l'image de départ.

3.2 Détection des points d'intérêts

La seconde étape de notre approche est celle de l'extraction de points d'intérêt. Pour celle-ci nous avons effectué une étude comparative entre différents détecteurs de coins: SUSAN, CSS et Harris. Le premier à la particularité de ne pas se baser sur une représentation mathématétique d'un coin mais sous la forme d'un masque de pixel, ce qui a pour avantage de choisir exactement le type de coins que l'on cherche et leur niveau de gris moyen. Mais s'il y a du bruit, la localisation devient approximative et la répétatbilité est moyenne. Le détecteur CSS se base sur des représentations à plusieurs échelles de contours (donc nécéssite Canny-Deriche) mais s'il est stable en terme de localisation, sa robustesse au bruit est dans notre cas, un peu faible. Enfin le détecteur de Harris s'est montré le plus stable dans nos tests sur les images sous-marines, selon le critère de répétabilité définit par Schmid et al. [15].

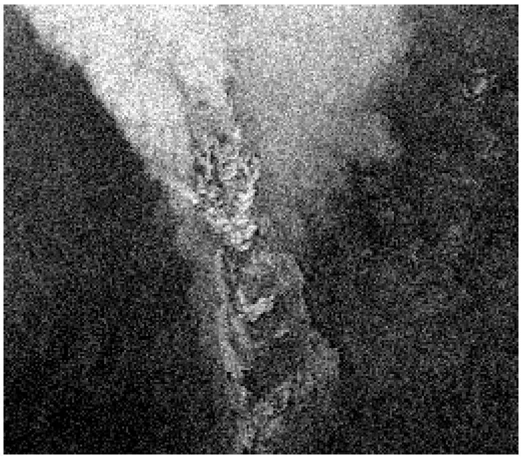


FIG. 2 – *Image bruitée de la fumerolle*

Nous ne présentons ici qu'un extrait des résultats obtenus lors de l'étude sur les différents détecteurs.

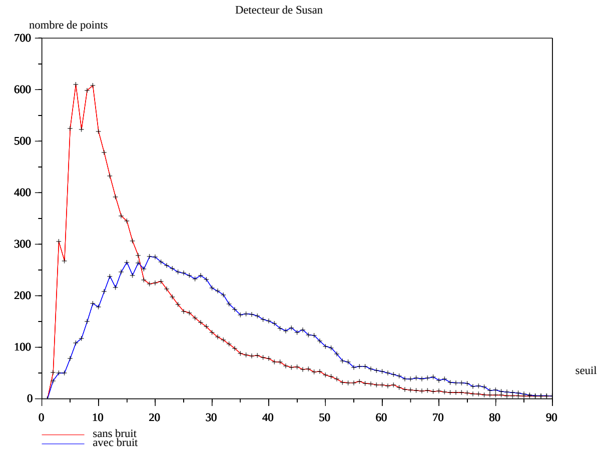


FIG. 3 – *Détecteur de Susan*

Sur la figure 3, on peut voir l'évolution du nombre de points en fonction du paramètre de seuillage pour le détecteur de SUSAN pour deux images: l'une bruitée (fig. 2) et l'autre non (fig. 1). On peut constater que le détecteur n'est pas très robuste au bruit dans l'image.

Les mêmes résultats sont présentés pour le détecteur de Harris sur la figure 4 et l'on peut constater que le détecteur est beaucoup plus stable et robuste. Les autres expériences (incluant aussi le détecteur CSS) ont confirmé ce fait et notre choix s'est donc porté vers le détecteur de Harris modifié par Schmid et implémenté avec les gaussiennes récursives de Deriche.

On rappelle brièvement le principe du détecteur de Harris qui est en fait un calcul de la fonction d'autocorrélation:

$$M = exp^{-\frac{x^2+y^2}{2\sigma^2}} \otimes \begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} \qquad (1)$$

où $I_x = \frac{\partial I}{\partial x}$ et $I_y = \frac{\partial I}{\partial y}$.

Puis il s'agit de déterminer les valeurs propres de

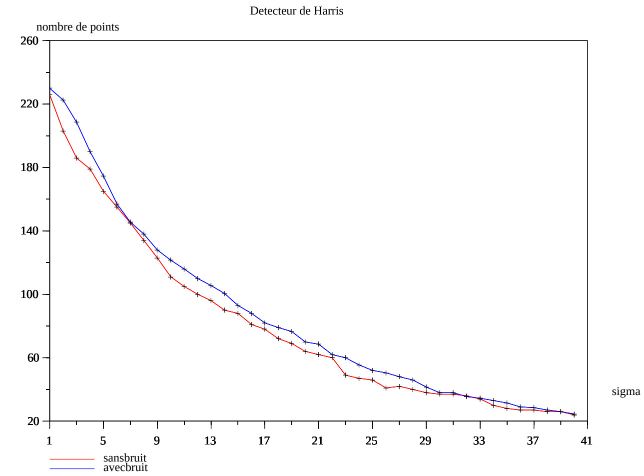


FIG. 4 – *Détecteur de Harris*

cette matrice, si elles sont grandes, alors il y a un coin. Harris et Stephens propose de calculer un critère: $h = Det(M) - k.Tr(M)^2$ et si celui-ci est supérieur à un seuil (en général, $h > 0$) alors il y a un coin. Habituellement, k est fixé à 0.04, ce qui correspond à une approximation au premier ordre d'une décomposition en valeurs singulières (méthode SVD) de la matrice M.

Afin de stabiliser le calcul du détecteur de Harris nous allons le robustifier [14] de la façon suivante: comme il faut dériver l'image, donc faire une lissage pour éviter les petites perturbations et effectuer la convolution, on a:
$\partial_i(g * f) = g * \partial_i f = \partial_i g * f$ où $*$ est l'opérateur de convolution et g la fonction gaussienne:

$$g(x,\sigma) = \frac{1}{2\pi\sigma^2} exp^{-\frac{x^2}{2\sigma^2}}$$

En résumé, pour une image donnée, nous avons construit sa pyramide associée et nous avons extrait un ensemble de points pour chaque niveau de celle-ci.

3.3 Appariement pyramidal

Cette partie est la plus importante du traitement car nous devons extraire des bons points. La notion de “bons points” est délicate à définir formellement. De façon générale, il faut que ces points détectés soient d'une part robustes aux bruits et aux lissages successifs et d'autre part, bien localisés.

L'extraction de points pour les différents niveaux ne se fait pas indépendamment les uns des autres. En effet, si nous réduisons la taille des images et que nous les lissons de plus en plus, le signal devient fortement dégradé et peu d'artéfacts vont subsister. A partir de cette observation, nous allons pouvoir trouver les points au niveau le plus élevé de la pyramide et les “propager” dans le reste de la pyramide vers l'image de départ de résolution maximale.

Cette propagation repose sur une propriété forte que nous qualifions d'injectivité entre deux points de deux niveaux successifs. En effet, un point détecté $p_{i,l}$ à un niveau l a également un antécédent $p_{j,l-1}$ au niveau inférieur $l-1$ mais n'est pas nécessairement exactement au même endroit du fait de la discrétisation et du lissage entre deux niveaux. On a donc la relation suivante: $p_{j,l-1} = f(p_{i,l})$

L'algorithme de recherche des correspondants entre le niveau l et $l-1$ est alors le suivant:

- Pour chaque point $p_k(x,y)$ du niveau k
  - créer une fenêtre de recherche de taille λ centrée en $(2x,2y)$
  - pour tout point $p_{k-1}(x',y')$contenu dans cette fenêtre, mettre à jour l'arbre d'appariement
- l'arbre obtenu comporte pour chaque noeud, ses fils ou et ses parents s'il en a

Le problème du réglage de ce paramètre λ se pose. En pratique, nous avons estimé qu'une fenêtre de recherche de taille 4x4 ou 5x5 suffisait amplement. En effet, lors de la construction de la pyramide, nous devons convoluer un point $p(x,y)$ avec une gaussienne $g(x,y,\sigma)$ puis nous sous-échantillonnons. Or, nous savons que: $\int_{-2\sigma}^{+2\sigma} g(t)dt = 0.99$. En clair, le support le gaussienne contribuant au résultat de la convolution avec f est de $[4\sigma]^2$ dans le cas 2d. Comme nous discrétisons avant de sous-échantillonner, nous prenons la partie entière supérieure:
$\lceil [g * f(x-2\sigma,y-\sigma), g * f(x+2\sigma,y+2\sigma)]^2 \rceil$, où $\lceil x \rceil$ désigne la partie entière suppérieure de x. Ce faisant, nous n'occasionnons pas de perte dans le calcul, car de toute façon les termes au delà de $(x,y) \pm 2\sigma$ ne sont pas pris en compte, car ils sont nuls.

Puis nous sous-échantillonnons: le support "actif" autour du point $p_{k-1}(x,y)$ au niveau $k-1$ devient alors:

$$\frac{1}{2} \lceil [g * f(x-2\sigma,y-\sigma), g * f(x+2\sigma,y+2\sigma)]^2 \rceil$$

Ainsi lors de l'appariement pyramidal et de la projection d'un point d'un niveau k à $k-1$, le support de recherche est donc deux fois celui du niveau inférieur, ce qui donne:

$$\lceil [g * f(x-2\sigma,y-\sigma), g * f(x+2\sigma,y+2\sigma)]^2 \rceil$$

En pratique σ varie entre 1 et 2.5, ce qui donne des zones de recherche qui vont jusqu'à 6x6. Une amélioration peut être de considérer la discrétisation optimale, à savoir choisir la partie entière la plus proche. Dans la réalité, cela n'influence que très peu le résultat.

En fin de compte, nous obtenons un ensemble d'arbres n-aires qui représentent les différents appariements dans la pyramide. Du fait de l'injectivité et de la nature des images, il n'est donc pas toujours possible de remonter au niveau supérieur. C'est cela qui nous permet de faire une classification sur la nature des points.

3.4 Classification

Comme dit précedemment, on trouve plusieurs ensembles de points à chaque niveau et surtout on sait quels sont ceux appariés ente deux niveaux. Il existe en fait un nombre limité de possibilités de configurations possibles. Par exemple, si un point peut être suivi tout au long de la pyramide, cela signifie que celui-ci est robuste aux différents lissages et le signal de départ a une forme spécifique d'artéfact très prononcé. De même, si un ensemble de points est détecté au niveau 0 et qu'aucun d'entre eux n'est apparié avec un point du niveau 1, cela indique que ces poins correspondent à du bruit.

On peut donc classifier les points de cette façon suivant ce tableau. Dans la colonne de gauche se trouve l'arbre d'appariement, dans celle de droite, la forme du signal en 1D. A partir de ces arbres il faut choisir quel est le point entre plusieurs qui sera prépondérant sur les autres.

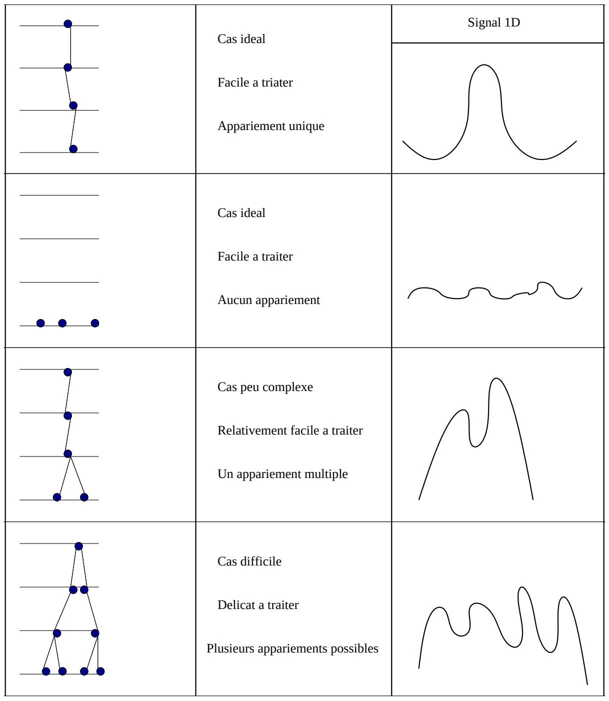


FIG. 5 – *Classification des points détectés*

Lorsque plusieurs choix sont possibles (lignes 3 et 4), on utilise la valeur du détecteur de Harris pour choisir quel sera le meilleur candidat lors de la "redescente" de l'appariement.

Cette étape est primordiale, et permet de créer des ensembles d'arbres et de sous arbres. On peut par exemple définir l'arbre primaire qui représente tous les points du niveau supérieur appariés de façon unique dans la pyramide.

4 Approche globale et Rigidité

Les points détectés par la méthode détaillée cidessus ne représentent que des informations locales dans l'image mais ne permettent pas de prendre en compte la notion de globalité ou de rigidité de l'objet.

Pour cela, nous utilisons les points détectés et les relions entre eux par une structure de graphe. Après expériences, nous avons opté pour une triangulation de Delaunay plutôt qu'un simple graphe connexe. En effet, le graphe ne correspond pas nécessairement à une bonne représentation de l'objet et sa complexité est grande (pour n points, on a $\frac{n(n-1)}{2}$ arcs). De plus, il est difficilement concevable de mettre des longueurs comme poids sur les arcs. Les maillages de Delaunay possèdent en outre plusieurs "bonnes" propriétés. Comme nous nous plaçons dans une optique de reconstruction 3d, nous devons avoir des points stables et éventuellement d'autres informations complémentaires. Le maillage de Delaunay représente localement la projection d'un maillage 3d d'une surface convexe. Bien évidemment il n'y pas bijection entre le maillage 2d et le maillage 3d.

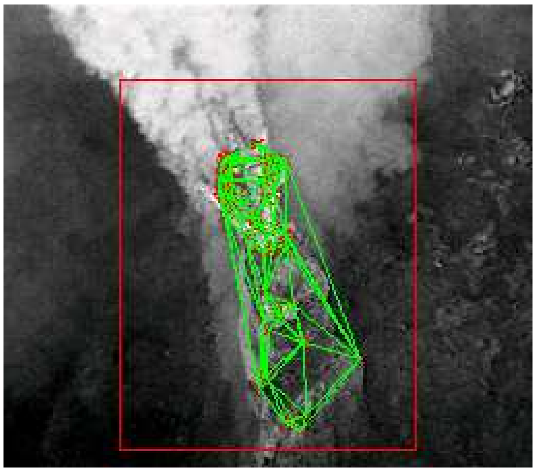


FIG. 6 – *Triangulation de Delaunay, image 0*

Sur la figure 6, on peut voir le résultat d'une triangulation de Delaunay sur un image de fumerolle. Bien que celui-ci soit de bonne qualité, il met aussi en évidence les limites de ce type de triangulation: en effet, on le constate, certains arcs ne correspondent à aucune réalité physique.

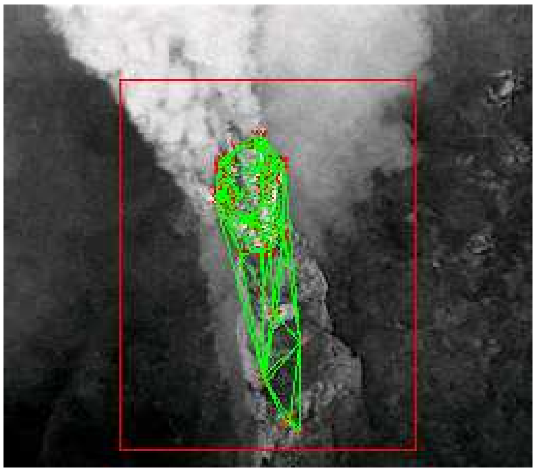


FIG. 7 – *Triangulation de Delaunay, image 4*

Néanmoins un des avantages de cette triangulation est, que dans le cas de deux images éloignées, la structure rigide de l'objet ne peut pas varier beaucoup et elle est préservée grâce au maillage de Delaunay comme on peut le voir sur la figure 7.

A partir de cette struture triangulée, on va pouvoir calculer des axes de symmétrie, des centres de masses, évaluer les déformations générales de l'objet. Un autre avantage de la triangulation de Delaunay est qu'il est possible de la contraindre, c'est-à-dire imposer certaines arêtes comme fixées pour la triangularisation.

5 Résultats

Nous avons validé notre approche sur plusieurs séries d'images et d'objets, comme la fumerolle sous marine ou encore des morceaux de liège, dont la structure rigide et la texture aléatoire et se rapproche fortement des objets sous-marins.

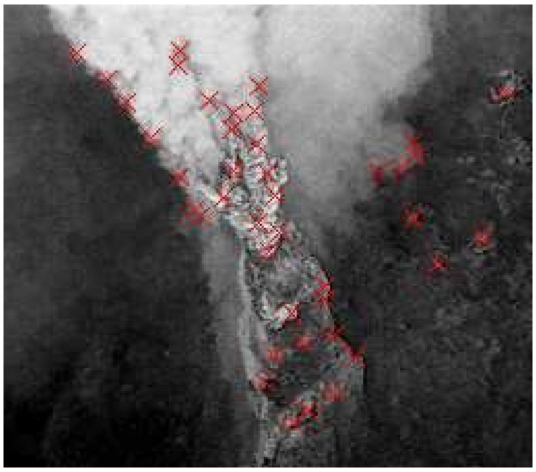


FIG. 8 – *Fumerolle*

Sur la figure 8, nous avons juste effectué un détecteur de Harris sur l'image entière sans se soucier de l'approche pyramidale. On le constate, si les points ont l'air corrects, il est difficile de les différencier.

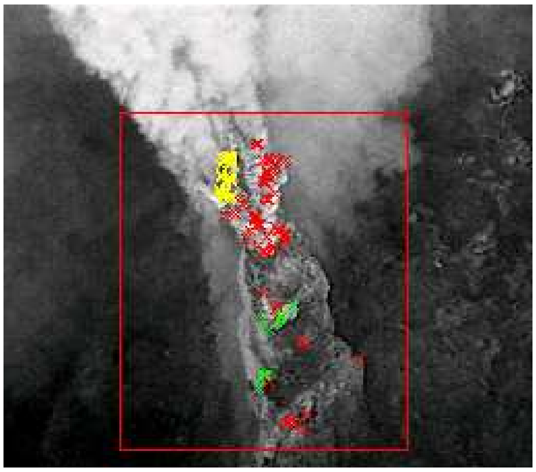
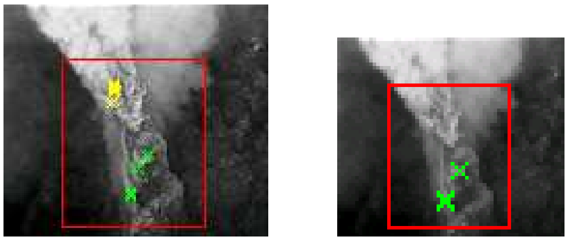


FIG. 9 – *Fumerolle, niveau 0*



(a) Niveau 1 (b) Niveau 2

FIG. 10 – *Points appariés avec l'algorithme pyramidal*

Sur les figures 9 et 10, on a utilisé l'approche pyramidale et et l'on distingue trois groupes de points. On peut alors identifier aisément les points stables, à savoir ceux qui sont présents dans tous les niveaux de la pyramide.

6 Conclusion et perspectives

Pour conclure, nous avons présenté dans cet article, une méthode permettant de robustifier l'extraction de points stables dans des images naturelles complexes non structurées et plus particulierement dans des images sous-marines. Celle-ci se base sur une utilisation du détecteur de Harris dans une pyramide d'images ayant pour but de sélectionner les meilleurs points à la fois aux lissages successifs et bien localisés. L'utilisation d'une triangulation de Delaunay permet d'inclure dans la représentation de l'image une notion de rigidité globale, par opposition aux points ou seule l'information locale est stockée.

A partir de cette représentation à la fois locale et globale, nous allons pouvoir aborder le problème de la reconstruction d'objets complexes. Cela se fera via une phase d'appariement robuste sur les points et sur le graphe. Le graphe construit nous permettra de calculer efficacement la matrice fondamentale grâce aux ensembles de points stables et d'avoir un modèle 3d projectif. l'estimation de la matrice fondamentale est un problème délicat et fera l'objet d'une attention particulière dans la suite du travail.

Une autre perspective de travail intéressante est l'estimation du mouvement de la caméra à partir des déformations du graphe. En effet, bien que le maillage 2d de Delaunay diffère entre deux images, il représente néanmoins toujours le même objet. La déformation entre les deux maillages peut certainement nous amener des informations quant au mouvement de la caméra et leur caractérisation formelle est un problème intéressant.

Références

[1] P. R. Beaudet. Rotationally Invariant Image Operators. In *Int. Joint Conf. on Pattern Recognition*, pages 579–583, 1978.

[2] J. Canny. A Computational Approach to Edge Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6):679–698, november 1986.

[3] R. Deriche. Using Canny's Criteria to Derive a Recursively Implemented Optimal Edge Detector. *International Journal of Computer Vision*, 1(2):167–187, may 1987.

[4] R. Deriche and G. Giraudon. Accurate Corner Detection: an Analytical Study. Technical Report RR 1420, INRIA, France, avril 1991.

[5] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Proc. of the 4th Alvey Vision Conference*, pages 147–151, 1988.

[6] M. Kass, A. Witkin, and D. Terzoppoulos. Snakes: Active Contour Models. *Int. Journal of Computer Vision*, 1:321–331, 1988.

[7] L. Kitchen and A. Rosenfeld. Gray-level corner detection. *Pattern Recognition Letters*, 1(2):95–102, december 1982.

[8] Pierre Kornprobst. *Contribution à la restauration d'images et à l'analyse de séquences: Approches Variationnelles et Solutions de Viscosité.* PhD thesis, Université de Nice, France, novembre 1998.

[9] M. Lhuillier. Joint View Triangulation for two Views. In *Orasis 99*, pages 151–158, Aussois, France, avril 1999.

[10] A. Lorette. *Analyse de texture par méthodes markoviennes et par morphologie mathématique: application à l'analyse des zones urbaines sur des images satellitales.* PhD thesis, Université de Nice - Sophia-Antipolis, France, septembre 1999.

[11] F. Mokhtarian and R. Suomela. Robust image corner detection through curvature scale space. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(12):1376–1381, december 1998.

[12] J. A. Noble. Finding Corners. *Image and Vision Computing*, 6(2):121–128, may 1988.

[13] C. Schmid and R. Mohr. Matching by local invariants. Technical Report RR 2644, INRIA, France, august 1995.

[14] C. Schmid and R. Mohr. Local Greyvalue Invariants for Image Retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19:530–534, may 1997.

[15] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and Evaluating Interest Points. In *Proc. of the 6th International Conference on Computer Vision*, pages 230–235, 1998.

[16] S. M. Smith and J. M. Brady. SUSAN - A New Approach to Low Level Image Processing. *Int. Journal of Computer Vision*, 23(1):45–78, may 1997.

[17] C. Tomasi and T. Kanade. Detection and Tracking of Point Features. Technical Report CS-91-132, Carnegie Mellon University, April 1991.