L'analyse des images dans les grands relevés d'astronomie visible/proche-infrarouge

Emmanuel Bertin[1]

[1]Institut d'Astrophysique de Paris
98bis, bd Arago, F-75014 Paris, France
bertin@iap.fr

Résumé – Le volume et la nature des données issues des grands relevés d'imagerie astronomique posent des problèmes spécifiques en matière de traitement et d'analyse. Je présente les réponses actuelles, mais aussi les questions en suspens concernant la détection et l'analyse morphologique des sources dans le domaine visible/proche-infrarouge.

Abstract – The amount and the nature of the data coming from large astronomical imaging surveys raises specific issues concerning image processing and analysis. I present both the current solutions and yet unanswered questions about source detection and morphological analysis in the visible/near-infrared domain.

1 Introduction : les images astronomiques

Les images obtenues aux longueurs d'ondes visibles et proche-infrarouge représentent, et de loin, le plus gros volume de données scientifiques archivées en astronomie. Ces images se caractérisent par un bruit de fond comprenant en proportions variables une composante poissonnienne (bruit de photons) et une composante gaussienne (électronique), auxquelles viennent s'ajouter des contaminations optiques plus ou moins corrélées avec la position des sources (réflexions parasites), des artefacts liés aux détecteurs (impacts de rayons cosmiques, pixels "chauds"), et enfin des "intrus" tels que les trainées de satellites. Sur ce fond, les astronomes cherchent à détecter et classer automatiquement les astres qui s'y trouvent. Ce court article passe en revue les difficultés et les solutions techniques mises en œuvre pour ces deux types de tâches, en particulier en France.

2 L'extraction de sources

On désigne par extraction de sources le processus consistant à générer un catalogue d'objets célestes contenant au minimum les mesures des positions et du flux des objets à partir d'une image. Le problème de l'extraction automatique des sources du ciel profond remonte au milieu des années 60, avec les premiers comptages informatisés de radiosources, et les premiers projets d'exploitation de l'énorme quantité d'informations s'accumulant depuis près de 80 ans sur les plaques photographiques [13]. De 1970 au milieu des années 1990, les machines à numériser les plaques photographiques furent les grands pourvoyeurs des catalogues automatisés à grande échelle du ciel visible.

Les techniques automatisées d'extraction de sources ont relativement peu évolué depuis, simplement parce que les performances des méthodes simples utilisées en astronomie remplissent les conditions requises par la plupart des programmes scientifiques. En effet, la tâche est relativement aisée, en comparaison d'autres problèmes d'analyse automatique d'image : nous avons ici affaire essentiellement à des objets intrinsèquement lumineux sur un fond sombre, et les effets de perspective ou d'ombre peuvent être ignorés. Il y a cependant quelques difficultés : les objets n'ont pas de bord net, qu'ils soient résolus ou non ; au sein d'une même image, les algorithmes doivent gérer une grande variété de structure et de taille des objets (jusqu'à un facteur 10000), et de rapports S/B (de 0 à près de 80dB). La réponse impulsionnelle est souvent variable et a une forte influence sur l'interprétation des images, en particulier l'identification des galaxies mal résolues (la population de sources largement dominante sur les trois-quarts du ciel). Heureusement, une mesure de la réponse impulsionnelle est naturellement disponible en de nombreux points du champ sous la forme d'images d'étoiles. Enfin, le volume de données à analyser (plusieurs Tpixels/an pour un relevé typique) impose des contraintes sur le débit des algorithmes, de l'ordre du Mpixel/s.

La majorité des algorithmes mis en oeuvre pour l'extraction de sources astronomiques faibles [51], [20], [21], [53], [18], [47], [3], [5], [11] fonctionnent sur le principe du seuillage d'images filtrées [43]. A quelques variantes près le schéma de traitement et d'analyse est le suivant :

1. Une distance angulaire maximale (de l'ordre de la minute d'arc) est fixée, déterminant l'échelle en pixels au-delà de laquelle une fluctuation sera considérée comme faisant partie du fond de ciel, et non une source à détecter. Cette échelle définit l'échelle de modélisation du fond de ciel (au moyen de splines par exemple).
2. Un filtrage linéaire adapté (*matched filter*), basé sur un profil de source légèrement plus large que la réponse impulsionnelle, et représentatif de la popula-

tion d'objets faibles, est appliqué à l'image soustraite du modèle de fond de ciel.

3. L'image filtrée est segmentée par seuillage (8-connexité) à différents seuils de valeur de pixel ; c'est donc la brillance de surface, et non le flux total qui détermine la détectabilité d'une source bien résolue.
4. Un arbre est construit à partir des composantes seuillées et une extraction de zones est décidée selon un critère de contraste local.
5. Les pixels voisins de chaque ensemble connexe identifié comme source, mais situés sous le seuil de détection, sont associés de manière probabiliste à chaque source, et les différentes mesures (position, flux, morphologie) sont effectuées (Fig. 1).

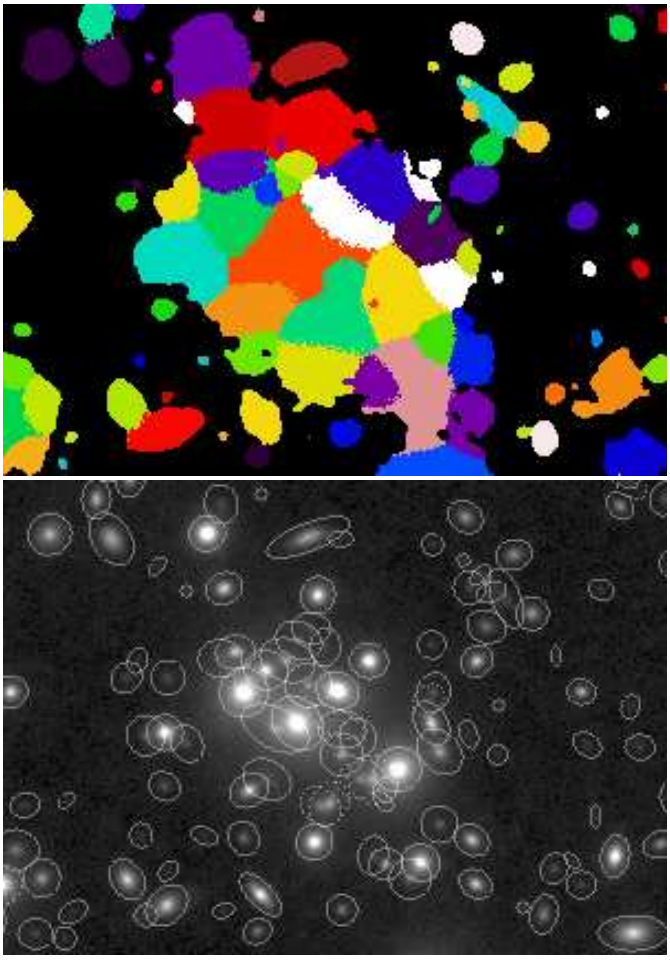

FIG. 1 – Cartes de segmentation initiale (*haut*) et d'identification des sources (*bas*) obtenues par le logiciel SExTRACTOR [5] sur un amas dense de galaxies faibles. Chaque ellipse identifie le centre et l'extension photométrique d'une galaxie détectée.

Le filtrage adapté est justifié par la stationnarité à moyenne échelle du bruit de fond, la contribution dominante pour la majorité des images du ciel profond. Les composantes basse-fréquence du bruit sont, elles, absorbées dans le modèle du fond. L'approche par multi-seuillage, bien que généralement rapide et performante, pose malgré-tout quelques problèmes pour les champs très encombrés d'étoiles (dans la voie lactée), dans lesquels une détection de pics est plus adaptée ([33], [16], [50], [29], [30]). Plus généralement, le multi-seuillage peut être avantageusement remplacé par un filtrage (isotropique) multi-échelle [42], [7], [22], [15], [8], [25], [49], [14], [31], notamment pour les images avec peu de photons par pixel, au prix cependant d'un impact parfois prohibitif sur le temps de calcul.

Convenablement paramétrées, les techniques "rustiques" que nous venons d'évoquer atteignent des taux de complétude et de fiabilité dépassant souvent 95% sur les simulations réalistes d'images astronomiques [28], une performance suffisante pour la grande majorité des applications scientifiques. Les quelques % de problèmes restants résident essentiellement dans des superpositions entre sources, ou avec des défauts optiques. Les résoudre demande un effort supplémentaire considérable, et probablement l'emploi de techniques d'apprentissage automatique, afin d'aboutir à un véritable modèle de vision [7]. Les quelques tentatives naïves faites à ce jour en ce domaine restent peu convaincantes [26], [1].

3 Classification morphologique

3.1 Séparation étoiles/galaxies

À la base, les sources détectées sur les images astronomiques peuvent être divisées en deux grandes classes : les sources ponctuelles (les étoiles) et les sources étendues (nébuleuses, et surtout galaxies). Le profil des sources ponctuelles isolées est celui de la réponse impulsionnelle locale ; distinguer automatiquement une étoile d'une galaxie est donc en principe chose aisée, et un classificateur optimal au sens bayesien peut être dérivé analytiquement [44], [53]. Malheureusement, les sources sont rarement parfaitement isolées ; compagnons d'étoiles doubles serrées, étoiles d'avant-plan et nébulosités d'arrière-plan contaminent largement les 2 classes. De plus, les images n'ont pas toujours un comportement linéaire en flux (saturations, images photographiques). Dans les années 90, l'introduction de techniques d'apprentissage automatique a permis de résoudre ce problème de manière simple et efficace : par logique floue [48], [27], arbres de décision [54], [19], [55], ou réseaux neuronaux supervisés [45], [4], [34], [5], [2], [1], [37] (Fig. 2).

3.2 Classification des galaxies

Les statistiques concernant les formes des galaxies sont au coeur de questions fondamentales de l'astrophysique moderne, en tant que marqueurs de la morphogenèse, de l'évolution des galaxies, des interactions, ou encore de distorsions produites par des lentilles gravitationnelles. Mais les catalogues de galaxies faibles construits jusqu'ici ne contiennent guère que la position, le flux, et les paramètres d'ellipse décrivant l'étendue apparente des sources, en omettant des informations telles que la présence de bras spiraux, d'un bulbe, d'une barre, d'un anneau, de queues de marée, de régions de formation stellaires,... Ces paramètres définissent en astronomie le *type* de la galaxie. Différents systèmes de classification existent, le plus connu

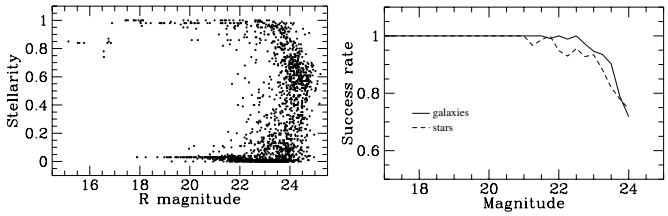

FIG. 2 – Classification automatique étoile/galaxie au moyen d'un MLP entrainé sur des simulations réalistes d'images astronomiques [5] . Les attributs utilisés sont des aires isophotales et l'intensité maximale des pixels. *Gauche* : valeur de sortie du réseau pour des sources réelles (0=galaxie, 1=étoile) en fonction de leur magnitude (grandeur logarithmique caractérisant le flux en astronomie, d'autant plus élevée que le flux est faible). *Droite* : taux de succès de classification pour les 2 classes, dérivé de simulations.

étant celui de Hubble [17], révisé par de Vaucouleurs [9].

L'analyse morphologique totalement automatique et fiable des centaines de millions de galaxies détectées dans les relevés du ciel est le grand défi actuel en matière d'analyse automatique des données astronomiques. Les premières tentatives de classification morphologique automatique à grande échelle remontent aux années 90 [48], [52], [35], [32]. Les algorithmes d'apprentissage (surtout rétropropagation sur perceptron multi-couche) atteignent des performances proches de celles d'experts humains [24]. Malheureusement, les classificateurs produits dans ces études ne sont pas transposables d'un relevé, voire d'une observation à l'autre. Les attributs sélectionnés sont en effet basés sur des mesures de contours très sensibles à la fois au bruit et à la réponse impulsionnelle.

Depuis quelques années, les recherches sont également motivées par les mesures fines à grande échelle des distorsions gravitationnelles faibles, et s'orientent vers un dispositif de classification plus indépendant des conditions de rapport S/B et de la réponse impulsionnelle des observations. L'immense majorité des galaxies détectées dans les grands relevés sont en effet faibles, mal résolues, et imagées dans des conditions de qualité variable. Une solution robuste est l'ajustement non-linéaire des paramètres d'un modèle analytique de galaxies convolué par la réponse impulsionnelle locale [46], [40], [38], [39]. Outre son coût en calcul, cette approche est malheureusement limitée à une décomposition bulbe+disque, sans prendre en compte d'autres caractéristiques morphologiques plus subtiles. Plus récemment ont été investiguées [6], [41], [23] des décompositions linéaires sur des bases de fonctions choisies de sorte qu'elles génèrent des vecteurs de coefficients corrigeables des variations de qualité d'image. Les implémentations actuelles se heurtent malheureusement encore à une gestion trop délicate des dégénérescences survenant sur les galaxies mal-résolues.

3.3 EFIGI

Dans ce contexte a été lancé en France fin-2004 le projet EFIGI [12], dans le cadre d'une Action Concertée Incitative CNRS de 3 ans, regroupant 2 laboratoires des STIC (LTCI et LRDE) et 5 laboratoires d'astronomie. Le but d'EFIGI est de créer pour la communauté un système performant de classification automatique d'images de galaxies, déclinable en *web-service* pour l'Observatoire Virtuel (voir F. Genova, cette conférence). Le projet contient des aspects de collecte et formatage d'échantillons (Fig. 3), de traitement du signal, d'apprentissage automatique, et de calcul distribué.

FIG. 3 – Extrait d'un échantillon d'images de galaxies bien résolues aux longueurs d'ondes visibles, issues du relevé SDSS [56] et compilées dans le cadre du projet EFIGI pour l'apprentissage de leur classification morphologique. En haut à droite de chaque vignette, est indiquée une série de codes décrivant l'aspect de la galaxie [10].

4 Conclusion

La détection et la classification automatiques des sources sont des passages obligés pour l'exploitation scientifique des grands relevés modernes d'imagerie astronomique. Ces deux activités représentent un terrain d'application stimulant pour les techniques de traitement du signal, en raison notamment des contraintes imposées par le volume des données et des exigences sévères de l'astrophysique sur le comportement statistique des algorithmes. Un projet tel qu'EFIGI illustre l'intérêt suscité par le rapprochement des communautés des STIC et de l'astronomie sur ces questions.

Références

[1] Andreon S., Gargiulo G., Longo G., Tagliaferri R., Capuano N., 2000, MNRAS **319**, 700

[2] Bazell D., Peng Y., 1998, ApJS **116**, 47

[3] Beard S.M., McGillivray H.T., Thanisch P.F., 1990, MNRAS **247**, 311

[4] Bertin E., 1994, Ap&SS **217**, 49

[5] Bertin E., Arnouts S., 1996, A&A **117**, 393

[6] Bertin E., Thion A., Mellier Y., van Waerbeke L., 2001, in *Gravitational Lensing : Recent Progress and Future Goals*, ASP Conf. Series **237**, 365

[7] Bijaoui A., Rué F., 1995, Signal Processing **46**, 229

[8] Damiani F., Maggio A., Micela G., Sciortino S., 1997, ApJ **483**, 350

[9] de Vaucouleurs G., de Vaucouleurs A., Corwin H.G., 1976, "Second Reference Catalogue of bright galaxies (RC2)", University of Texas Press, Austin

[10] de Vaucouleurs G., de Vaucouleurs A., Corwin H.G., Buta R.J., Paturel G., Fouqué P., 1991, "Third Reference Catalogue of bright galaxies (RC3)", Springer-Verlag, New York

[11] Drory N., 2003, A&A **397**, 371

[12] EFIGI : Extraction de Formes Idéalisées de Galaxies en Imagerie, `http ://www.efigi.org`

[13] Fellgett P.B., 1970, Optics Technology **2**, 61

[14] Freeman P.E., Kashyap V., Rosner R., Lamb D.Q., 2002, ApJS **138**, 185

[15] Grebenev S.A., Forman W., Jones C., Murray S., 1995, ApJ **445**, 607

[16] Herzog A.D., Illingworth G., 1977, ApJS **33**, 55

[17] Hubble E.P., 1926, ApJ **64**, 321

[18] Irwin M.J., 1985, MNRAS **214**, 575

[19] Jarrett T.H., Chester T., Cutri R., Schneider S., Skrutskie M., Huchra J.P., 2000, AJ **119**, 2498

[20] Jarvis J.F., Tyson J.A., 1979, in "Instrumentation in Astronomy III", SPIE **172**, 422

[21] Jarvis J.F., Tyson J.A., 1981, AJ **86**, 476

[22] Kaiser N., Squires G., Broadhurst T., 1995, ApJ **449**, 460

[23] Kelly B.C., McKay T.A., 2004, AJ **127**, 625

[24] Lahav O., Naim A., Buta R.J., Corwin H.G., de Vaucouleurs G., Dressler A., Huchra J.P., van den Bergh S., Raychaudhury S., Sodré Jr. L., Storrie-Lombardi M.C., 1995, Science **267**, 859

[25] Lazzati D., Campana S., Rosati P., Panzera M.R., Tagliaferri, G., 1999, ApJ **524**, 414

[26] Mähönen P., Hakala P.J., 1995, ApJ **452**, 77

[27] Mähönen P., Frantti T., 2000, ApJ **541**, 261

[28] McCracken H.J., Le Fèvre O., Foucaud S., Lilly S.J., Crampton D., Mellier Y., 2001, A&A **376**, 756

[29] Mighell K.J., 1989, MNRAS **238**, 807

[30] Mighell K.J., 1999, in "Astronomical Data Analysis Software and Systems VIII", ASP Conf. Series **172**, 317

[31] Moretti A., Lazzati D., Campana S., Tagliaferri G., 2002, ApJ **570**, 502

[32] Naim A., Lahav O., Sodré L. Jr, Storrie-Lombardi M.C., 1995, MNRAS **275**, 567

[33] Newell B., O'Neil,Jr E.J., 1977, PASP **89**, 925

[34] Odewahn S.C., Stockwell E.B., Pennington R.L., Humphreys R.M., Zumach W.A., 1992, AJ **103**, 318

[35] Odewahn S.C., 1995, PASP **107**, 770

[36] Odewahn S.C., 1997, in "Applications of Digital Image Processing XX", SPIE **3164**, 110

[37] Odewahn S.C., de Carvalho R.R., Gal R.R., Djorgovski S.G., Brunner R., Mahabal A., Lopes P.A.A., Kohl Moreira J.L., Stalder B., 2004, AJ **128**, 3092

[38] Peng C.Y., Ho L.C., Impey C.D., Rix H.-W., 2002, AJ **124**, 266

[39] Pignatelli E., Fasano G., Cassata P., 2005, A&A, à paraître

[40] Ratnatunga K.U., Griffiths R.E., Ostrander E., 1999, AJ **118**, 86

[41] Refregier A., 2003, MNRAS **338**, 35

[42] Rosati P., Burg R., Giacconi R., 1994, in "The Soft X-ray Cosmos", Eds. E.M. Schlegel,R Petre, AIP Conf. Proceedings **313**, 260

[43] Rosenfeld A., 1969, in "Picture Processing by Computer", Academic Press, New-York, 127

[44] Sebok W.L., 1979, AJ **84** 1526

[45] Serra-Ricart M., Gaitan V.,Delgado S., Perez-Fournon I., 1992, in "Astronomical Data Analysis Software and Systems I", ASP Conf. Series **25**, 254

[46] Simard L., 1998, in *Astronomical Data Analysis Software and Systems VII*, ASP Conf. Series **145**, 108

[47] Slezak E., Bijaoui A., Mars G., 1988, A&A **201**, 9

[48] Spiekermann G., 1992, AJ **103**, 2102

[49] Starck J.-L., Bijaoui A., Valtchanov I., Murtagh F., 2000, A&AS **147**, 139

[50] Stetson P.B., 1987, PASP **99**, 191

[51] Stobie R.S., Smith G.M., Lutz R.K., Martin R., 1979, in "International Workshop on Image Processing in Astronomy", Eds. G. Sedmak, M. Capaccioli, R.J. Allen., Osservatorio Astronomico di Trieste, Trieste, 48

[52] Storrie-Lombardi M. C., Lahav O., Sodré L. Jr, Storrie-Lombardi L.J., 1992, MNRAS **259**, 8

[53] Valdes F., 1982, in "Instrumentation in Astronomy IV", SPIE **331**, 465

[54] Weir N., Fayyad U.M., Djorgovski S., 1995, AJ **109**, 2401

[55] White R.L., 2000. in "Astronomical Data Analysis Software and Systems IX", ASP Conf. Series **216**, 577

[56] York D.G. et al., 2000, AJ **120**, 157