Evaluation de la segmentation d'images : état de l'art, nouveaux indices et comparaison

Sylvie Philipp-Foliguet*, Laurent Guigues**
*ETIS, UMR CNRS 8051
6 avenue du Ponceau, 95014 Cergy Cedex
philipp@ensea.fr
**CREATIS, UMR CNRS 5515, Inserm U 630
INSA, 7 rue Jean Capelle, 69621 Villeurbanne Cedex
laurent.guigues@creatis.insa-lyon.fr

Résumé : Cet article s'intéresse à l'évaluation de méthodes de segmentation d'images en régions. Nous commençons par un état de l'art des critères d'évaluation, en séparant le cas où l'on dispose d'une segmentation de référence de celui où l'on n'en dispose pas. Nous appuyant sur une analyse des principaux critères existants, nous proposons alors de nouveaux critères pour le cas où l'on ne dispose pas de référence. Ces critères, basés sur une formulation énergétique, prennent en compte à la fois la complexité de la segmentation et la fidélité aux données initiales d'un modèle sous-jacent. L'intérêt est de ne pas figer le niveau de détail attendu mais de le laisser au choix de l'utilisateur en fonction de son but. Les critères proposés sont donc des critères d'évaluation multi-échelle. Dans un premier temps, nous comparons les performances expérimentales de différentes formulations énergétiques possibles. Dans un second temps, ayant fait un choix d'énergie particulier, nous comparons les performances de notre nouveau critère avec celles des principaux critères existants.

Mots-clé : Image, segmentation, évaluation, énergie, échelle, critère

Abstract : This paper deals with evaluation of image segmentation methods. We start with a state-of-the art of the evaluation criteria, involving a reference segmentation or not. Based on an analysis of the main exiting criteria, we propose new criteria, when no ground-truth is available. These criteria, based on an energetical formalism, take into account both the complexity of the segmented image and the goodness-of-fit of an underlying model with the initial data. The main interest is not to fix the expected level of resolution but to leave it to the user's choice, according to his purpose. These evaluation criteria are thus multi-scale criteria. Various forms of energy formulation are experimentally compared. Then after having choosen a particular energy, the performances of this new criterion is compared to the main existing criteria.

Keywords: Image, evaluation, segmentation, criterion, energy, scale

1 Introduction

Devant le foisonnement de méthodes développées depuis plusieurs décennies pour la segmentation des images, le problème de l'évaluation est devenu crucial. Disposer de méthodes d'évaluation de résultats est nécessaire :

- aux chercheurs pour comparer un nouvel algorithme à ceux préexistants
- aux utilisateurs pour choisir un algorithme et régler ses paramètres en fonction du problème à résoudre.

Les critères d'évaluation quantitative peuvent être groupés en deux classes, selon que l'on possède ou non une "*vérité-terrain*" qui constitue une *segmentation de référence*. Celle-ci est directement accessible dans le cas d'images de synthèse, mais elle doit être construite "à la main" par un expert du domaine de l'application dans le cas d'images réelles : tracés effectués par des médecins, des géographes, etc... à l'aide d'outils informatiques de dessin.

Si l'on veut comparer de manière objective les méthodes, il est plus simple d'utiliser des images de synthèse, pour lesquelles une "vérité" est parfaitement connue, à savoir la segmentation qui a servi à synthétiser l'image. L'inconvénient d'une telle démarche est que ces images ne représentent pas toutes les situations possibles d'une utilisation réelle.

Bien que l'évaluation sur images réelles soit certainement plus réaliste, elle pose d'autres difficultés, la principale étant qu'il n'existe généralement pas de solution unique à la division d'une image en régions "pertinentes". La "pertinence" d'une région est en effet une notion éminemment dépendante de l'application : qui cherche à suivre les individus dans une vidéo sera intéressé par le détourage des silhouettes, qui s'intéresse à la mode voudra isoler les vêtements des personnages, etc. Néanmoins, deux segmentations humaines d'une même image tendent à être cohérentes dans le sens où elles sont des raffinements mutuels l'une de l'autre [2] : certaines régions d'une segmentation constituent une sur-segmentation de certaines régions de l'autre et inversement. Autrement dit, la principale différence entre deux segmentations humaines d'une même image est une différence de niveau de détail (nous en verrons des illustrations à la section 4).

Yu et Shi [28] ont récemment proposé une classification des méthodes de segmentation en deux grandes catégories : d'une part une approche dite "discriminative" qui, dans la droite ligne des méthodes de classification non supervisée, envisage la segmentation comme un problème de regroupement des pixels en classes compactes et bien séparées et d'autre part une approche dite "générative", de type "problème inverse", qui envisage la segmentation comme un problème de recherche d'un "modèle générateur" des données (generative model). Or le problème de l'évaluation de la segmentation sans référence est extrêmement proche du problème de segmentation lui-même. Nous verrons que les critères d'évaluation sans référence suivent cette même classification en deux grands groupes, l'un basé sur l'approche "discriminative" aboutit à des

mesures de type contraste et l'autre basé sur l'aproche générative, conduit à une modélisation par morceaux de l'image.

C'est cette deuxième approche que nous avons exploitée en nous appuyant sur une formulation énergétique multi-échelle de la segmentation [9, 11]. Nous proposons des critères d'évaluation qui prennent en compte explicitement le niveau de détail auquel l'utilisateur souhaite aboutir. Ces critères permettent d'ordonner les différentes segmentations concurrentes en fonction d'un paramètre d'échelle et de rejeter les segmentations qui ne sont pertinentes à aucune échelle.

La section 2 commence par faire un état de l'art des critères d'évaluation existant dans la littérature, en séparant le cas où l'on dispose d'une vérité-terrain de celui où l'on n'en dispose pas.

En s'appuyant sur une critique des critères existants, la section 3 développe alors les nouveaux critères proposés pour l'évaluation sans référence.

La section 4 présente finalement des résultats expérimentaux. Dans un premier temps, nous avons évalué différentes énergies possibles pour l'évaluation multi-échelle. Dans un second temps, nous avons comparé les nouveaux critères proposés aux principaux critères existants, à la fois sur des résultats de segmentation automatique et sur des résultats de segmentations manuelles.

2 Etat de l'art

Quand on dispose d'une vérité-terrain, l'évaluation des segmentations s'effectue à l'aide de critères comparant chaque segmentation avec l'image de référence. On peut ainsi ordonner les segmentations.

En l'absence de vérité-terrain, il faudra employer des critères quantitatifs absolus ou des calculs de cohérence entre les différents résultats de segmentation.

Dans la suite de l'article, une image I est définie sur un ensemble de sites X représentant les coordonnées spatiales des pixels (ligne, colonne) et une fonction f à valeurs dans un ensemble Z. Par exemple f pourra être l'intensité pour les images à niveaux de gris (dans ce cas Z est un sous-ensemble de N) et la couleur dans l'un des espaces colorimétriques pour les images couleurs (dans ce cas Z est un sous-ensemble de N^3).

On notera R une segmentation à évaluer, c'est donc une partition de X en régions notées R_i, $i = 1, ..., N$ vérifiant $R_i \cap R_j = \emptyset$ et $\bigcup_{i=1}^{N} R_i = X$. On note A le nombre de pixels de l'image et A_i le nombre de pixels de la région R_i. On a donc : $A = \sum_{i=1}^{N} A_i = card\, X$.

2.1 Avec segmentation de référence

Dans cette partie, nous disposons d'une segmentation de référence notée V, dont les régions sont notées V_i, $i = 1, ..., M$.

2.1.1 Mesure de Vinet [25]

La mesure de Vinet s'appuie sur un appariement biunivoque entre les régions des deux segmentations à comparer. Pour tout couple de régions (V_i, R_j), on définit leur recouvrement par $t_{ij} = card(V_i \cap R_j)$. Un couplage de poids maximal du graphe bipartite (V, R, t) fournit alors un appariement biunivoque optimal entre régions des deux segmentations au sens de la somme des recouvrements des parties appariées. Soit K le nombre de couples obtenus et $C_1 \dots C_K$ les recouvrements de chacun de ces couples. $\frac{1}{A} \times \sum_{k=1}^{K} C_k$ représente alors le poids total du couplage normalisé par la surface de l'image.

La mesure de dissimilarité de Vinet est alors : $1 - \frac{1}{A} \times \sum_{k=1}^{K} C_k$.

Bien que le couplage ne soit pas nécessairement unique cette mesure est une distance.

Une approximation de la mesure de Vinet peut être obtenue efficacement par un algorithme glouton qui consiste à coupler itérativement les deux régions de recouvrement maximal.

Cette mesure de dissimilarité a été utilisée dans [5] pour comparer des segmentations sur des images synthétiques monochromes comportant différents bruits et textures.

Des mesures généralisant la distance de Vinet à des appariements multivoques ont été proposés dans [8] et [2].

2.1.2 Mesure de cohérence entre segmentations de Martin [18]

Définie par D. R. Martin pour évaluer la cohérence entre deux segmentations manuelles (voir Fig. 8) d'une même image, cette mesure peut être utilisée pour comparer deux segmentations l'une de référence, l'autre obtenue par un algorithme.

Elle est basée sur deux erreurs calculées en chaque pixel : une erreur de V par rapport à R et une erreur de R par rapport à V. Si le pixel s appartient à la région V_j dans la vérité-terrain et à la région R_i dans l'image résultat, ces erreurs valent : $E(s) = \frac{card(V_j \setminus R_i)}{card(V_j)}$ et $E'(s) = \frac{card(R_i \setminus V_j)}{card(R_i)}$. $E(s)$ vaut 0 si V_j est un sous-ensemble de R_i et vaut 1 si l'intersection des deux régions est réduite au pixel s.

La dissimilarité entre segmentation résultat et segmentation de référence se mesure alors par l'erreur locale de cohérence

$$LCE\,(R, V) = \frac{1}{A} \sum_{s} \min\{E(s), E'(s)\}$$

ou par l'erreur globale de cohérence

$$GCE\,(I, V) = \frac{1}{A} \min\left\{ \sum_{s} E(s), \sum_{s} E'(s) \right\}$$

Cette dernière mesure est plus sévère que la première et a le défaut de favoriser une sur-segmentation (ou une sous-segmentation) de toute l'image par rapport à un mélange des deux situations (sur et sous-segmentation selon les zones de l'image).

2.1.3 Position des pixels mal segmentés : mesure de Yasnoff et al. [27]

Compter simplement le nombre de pixels mal segmentés est insuffisant, il faut aussi tenir compte de la position de ces pixels en utilisant par exemple la distance entre un pixel mal segmenté et la région à laquelle il appartient dans la référence.

La mesure de Yasnoff et al. s'écrit :

$$\frac{100}{A} \times \sqrt{\sum_{s} d^2(s)}$$

où la sommation s'effectue sur les pixels mal segmentés et d est la distance au pixel le plus proche de la région à laquelle il appartient.

Cet indice a été utilisé dans [29] pour comparer des méthodes de seuillage.

2.1.4 Distance de Baddeley [26]

Cette distance prend en compte non seulement la position du site s dans l'image mais également son intensité.

Soit une image d'intensité $f : X \longrightarrow Z \subset N$.

Le sous-graphe de f est : $\Gamma_f = \{(s, z),\ s \in X,\ z \in Z,\ \text{et}\ \ z \leq f(s)\}$

La distance entre un couple de $X \times Z$ et le sous-graphe de f est définie par : $d_B((s, z), \Gamma_f) = inf\{d((s, z), (s', z')), (s', z') \in \Gamma_f\}$

Cette distance est seuillée pour ne pas rechercher trop loin de z le minimum.

Et la distance entre deux images f et g est finalement :

$$\left(\frac{1}{A^2} \sum_{(s,z)} |d_B((s, z), \Gamma_f) \ - \ d_B((s, z), \Gamma_g)|^p \right)^{\frac{1}{p}} \quad \text{avec} \quad p \geq 1.$$

Cette distance est souvent citée. Elle peut s'utiliser en prenant pour d une distance de chanfrein au lieu de la distance euclidienne. Elle a été récemment étendue aux images couleur [6].

2.2 Sans segmentation de référence

De nombreux critères ont été proposés, cherchant à quantifier la qualité ou la lisibilité de l'image. Suivant la classification des méthodes de segmentation proposée dans [28], on peut classer ces critères en deux grandes catégories : les critères de "contraste" et les critères d' "adéquation à un modèle". Les premiers recherchent une variabilité inter-région, alors que les seconds recherchent une uniformité en intensité ou en couleur à l'intérieur des régions.

Parmi les critères de contraste, nous présentons ci-dessous celui de Levine et Nazif et celui de Zeboudj, ainsi que le critère de Rosenberger.

Parmi les seconds, nous présentons le critère d'uniformité de Levine et Nazif, le critère de Liu et Yang, et celui de Borsotti *et al.*

Si f représente un attribut du pixel (en général son intensité ou sa couleur), on notera m_i (resp. σ_i) la moyenne (resp. l'écart-type) de f dans la région R_i.

2.2.1 Contraste inter-région de Levine et Nazif [16]

Soit $c_{ij} = \frac{|m_i - m_j|}{m_i + m_j}$ le contraste entre deux régions adjacentes R_i et R_j.

Le contraste de la région R_i est :

$c_i = \sum_{R_j} p_{ij} c_{ij}$ où les R_j sont les régions adjacentes à R_i et

$p_{ij} = \frac{l_{ij}}{l_i}$ est le rapport longueur de la frontière commune entre R_i et R_j sur le périmètre de R_i.

Le contraste global est alors : $\frac{\sum_{R_i} w_i c_i}{\sum_{R_i} w_i}$

w_i est un poids associé à chaque région, qui peut être l'aire de la région.

Ce critère, utilisé par Zhang [29] sans pondération s'est révélé absolument non discriminant !

2.2.2 Contraste de Zeboudj [5]

Cet indice prend en compte le contraste intérieur et le contraste extérieur aux régions, mesurés sur un voisinage $W(s)$ du pixel s.

Soit $c(s,t) = \frac{|f(s) - f(t)|}{L-1}$ le contraste entre deux pixels s et t, avec f représentant l'intensité et L le maximum des intensités.

Le contraste intérieur d'une région R_i est :

$I_i = \frac{1}{A_i} \sum_{s \in R_i} \max\{c(s,t), t \in W(s) \cap R_i\}$

Le contraste extérieur d'une région R_i est :

$E_i = \frac{1}{l_i} \sum_{s \in F_i} \max\{c(s,t), t \in W(s), t \notin R_i\}$ où F_i est la frontière de R_i et l_i la longueur de F_i.

Le contraste de R_i est :

$$C(R_i) = \begin{cases} 1 - \frac{I_i}{E_i} & \text{si} \quad 0 < I_i < E_i \\ E_i & \text{si} \quad I_i = 0 \\ 0 & \text{sinon} \end{cases} \tag{1}$$

Le contraste global est enfin : $\frac{1}{A} \sum_i A_i . C(R_i)$.

Cet indice a été utilisé dans [5] pour comparer des segmentations en régions sur des images réelles et de synthèse (cf. 2.5.2). Cet indice n'est pas adapté aux images trop bruitées ou texturées.

2.2.3 Critère de Rosenberger [3]

Pour résoudre le problème des images monochromes contenant des textures, Rosenberger commence par caractériser chaque région en région texturée ou uniforme, grêce à un calcul d'uniformité des niveaux de gris basé sur les matrices de cooccurrences. Il calcule ensuite la disparité intra-région, notée $\underline{D}$ et la disparité inter-région, notée $\overline{D}$. La première correspond à l'écart-type des intensités pour une région uniforme et à un ensemble d'attributs de texture pour une région texturée. $\overline{D}$ est égale à la différence des moyennes pour deux régions uniformes, à la distance euclidienne entre attributs de texture pour deux régions texturées et à 1 pour une région texturée et une région uniforme.

La disparité intra-région globale est égale à la moyenne pondérée des disparités calculées pour chaque région :

$\underline{D} = \frac{1}{N} \sum_i \frac{A_i}{A} \underline{D}_i$

et de même pour la disparité inter-région globale.

Finalement le critère de Rosenberger est égal à : $\frac{\overline{D} - \underline{D}}{2}$

2.2.4 Critère d'uniformité intra-région de Levine et Nazif [16]

Ce critère simple est basé sur la somme des variances des régions. Il doit donc être faible.

$$\sum_i \sum_{s \in R_i} \left[f(s) - \frac{1}{A_i} \sum_{s \in R_i} f(s) \right]^2 = \sum_i \frac{\sigma_i^2}{C} \tag{2}$$

f peut être l'intensité du pixel s ou tout autre attribut (couleur, texture).

C est un facteur de normalisation, égal à la variance maximale :

$\sigma^2_{\max} = \frac{(f_{\max} - f_{\min})^2}{2}$

On peut également pondérer chaque région par son nombre de pixels.

L'avantage de ce critère est d'être facilement mis à jour dans les opérations de fusion ou de division des régions.

Il a été utilisé pour comparer des méthodes de seuillage dans [22] (voir § **??**) et des méthodes de segmentation [29], où il s'est révélé peu discriminant. Il fait partie des critères testés au paragraphe 4.2.

2.2.5 Mesure de dissimilarité de Liu et Yang [17]

Ce critère est basé sur le nombre de régions, l'aire des régions et la couleur moyenne, dans l'espace RGB :

$$\frac{1}{1000 \times A} \sqrt{N} \sum_{i=1}^{N} \frac{e_i^2}{\sqrt{A_i}} \tag{3}$$

où e_i est la somme des distances euclidiennes entre les vecteurs couleur des pixels de la région R_i et le vecteur couleur attribué à la région R_i dans l'image segmentée (en général la moyenne des couleurs de la région).

Le critère doit être faible. Les termes $\sqrt{N}$ au numérateur et $\sqrt{A_i}$ au dénominateur de l'expression pénalisent la sur-segmentation. Il est assez proche de celui de Levine et Nazif, le calcul des écarts à la moyenne est légèrement différent, ainsi que la normalisation, mais l'idée générale reste la même.

Cette mesure de dissimilarité a été utiliséepar leurs auteurs [17] pour rechercher le meilleur espace colorimétrique pour la segmentation. Malheureusement, aucun espace ne s'est dégagé comme meilleur pour tout type d'image. Nous l'utiliserons dans nos tests comparatifs du paragraphe 4.2.

2.2.6 Critère de Borsotti [1]

La mesure de dissimilarité de Liu et Yang pénalise les segmentations contenant trop de régions ou avec des régions non homogènes en couleur.

Borsotti *et al.* ont proposé de l'améliorer par :

$$\frac{1}{10000 \times A}\sqrt{N}\sum_{i=1}^{N}\left(\frac{e_i^2}{1+\log A_i}+\frac{N(A_i)^2}{A_i^2}\right) \tag{4}$$

où $N(A_i)$ est le nombre de régions ayant une aire égale à A_i.

Ce critère doit aussi être faible. Le premier terme de la somme favorise les régions homogènes, comme le critère de Liu et Yang. Le deuxième terme a une valeur élevée quand il y a beaucoup de petites régions, ce qui pénalise les images sur-segmentées en beaucoup de régions de même taille.

D'après les auteurs [1], ce critère a fourni sur 500 images un classement des segmentations plus conformes à notre appréciation visuelle que le critère de Liu et Yang. Il sera comparé aux autres critères dans le paragraphe 4.2.

2.3 Avec calcul d'attributs

La segmentation étant rarement l'étape ultime du traitement, celle-ci peut être évaluée par les traitements ultérieurs. Il existe quantité de mesures que l'on peut effectuer sur un objet ou un ensemble d'objets extraits de l'image.

Si l'on dispose d'un ensemble d'exemples des objets à extraire, on peut évaluer la segmentation d'un objet par les méthodes employées en classification : distance aux k plus proches voisins dans la classe, distance à un prototype (centre de gravité par exemple), etc. Les attributs de régions les plus simples sont les critères géométriques tels que l'aire, le périmètre, et des facteurs de forme comme la circularité, la courbure moyenne, etc.

On peut aller plus loin et mesurer la qualité de la segmentation grâce à la qualité de la classification des objets extraits. Les classifieurs flous offrent ainsi l'avantage, par le degré d'appartenance aux classes qu'ils fournissent de donner une évaluation de la segmentation [12].

2.4 Comparaison de critères

Y. J. Zhang [29] a comparé certains de ces critères pour évaluer des méthodes de seuillage binaire (recherche d'un seuil global sur l'image). Les tests ont été

effectués sur des images de synthèse comportant deux régions, bruitées par un bruit additif gaussien de moyenne nulle. 5 images sont ainsi formées avec 5 écarts-type de bruit différents. Les résultats sont ensuite moyennés sur les 5 images. Le but est de déterminer pour 14 valeurs de seuil, quel est le critère d'évaluation qui discrimine le mieux les résultats (d'une valeur de seuil à l'autre). 5 critères ont été testés, 3 utilisent la vérité-terrain.

1. mesure d'attribut de régions, en l'occurrence l'aire des régions (sur un seuillage et seulement deux régions, on se doute que c'est très discriminant !)
2. probabilité de mauvaise classification :$P(O)\times P(B|O)+P(B)\times P(O|B)$,où O est l'objet et B est le fond
3. mesure de Yasnoff et al
4. mesure d'uniformité intra-région de Levine et Nazif
5. le contraste inter-région de Levine et Nazif

Les trois meilleurs résultats sont obtenus par comparaison avec la vérité-terrain (les trois premiers critères).

Dans Laurent et al. [14] la plupart des critères cités dans ce paragraphe ont été comparés sur une base de 100 images synthétiques avec ou sans texture. Les critères de Zéboudj et le contraste inter-région de Levine et Nazif se sont révélés les plus performants pour tous les types d'images sauf celles contenant exclusivement des zones texturées pour lesquelles le contraste de Rosenberger est le plus performant. Il est intéressant de noter que bien qu'étant normalisés, certains critères ont une amplitude très faible (Borsotti et intra-région de Levine), et que d'autres critères attribuent une note assez médiocre à la segmentation idéale.

2.5 Emploi des techniques d'évaluation

2.5.1 Fiabilité de la vérité-terrain

L'image constituant la segmentation de référence, ou vérité-terrain est soit parfaitement connue s'il s'agit d'une image de synthèse, soit issue d'une segmentation manuelle effectuée par un ou plusieurs experts du domaine d'application.

Le problème de la variabilité du tracé manuel n'est pas négligeable. Il a été étudié par Chalana et Kim dans le cas d'images médicales [4]. Pour le problème d'un seul contour, en l'occurrence fermé, ils calculent un indice de variabilité inter-expert ainsi qu'un contour moyen, qui constituera le contour de référence. Les résultats de segmentation sont alors comparés à ce contour moyen, en liaison avec la variabilité inter-expert.

Un gros travail pour élaborer une base d'images de vérité-terrain a été mené par D. R. Martin [18]. 1020 images couleurs issues de la base Correl ont été segmentées manuellement par une trentaine de personnes, fournissant 11 595

segmentations disponibles sur Internet (cf. Figs. 6 et 8). Nous les utiliserons pour valider notre approche dans le paragraphe 4.2

2.5.2 Evaluation de segmentation en régions

Sur images de synthèse Dans [5], des images de synthèse contenant différents bruits additifs et des textures ont été employées pour comparer six méthodes de segmentation en régions. Les critères sont la mesure de Vinet (cf. § 2.1.1) et le contraste de Zeboudj (cf. § 2.2.2). Les deux indices ne sont pas toujours cohérents, mais globalement, les méthodes markoviennes, et notamment la méthode supervisée de relaxation fournissent les meilleurs résultats sur les images bruitées. Par contre, seule la classification multi-attributs (qui arrive dernière sur les images de bruit) parvient à extraire quelques régions des images texturées.

Sans référence Les mêmes algorithmes que précédemment plus un septième (extraction de contour, suivi d'une fermeture des contours) ont été comparés sur la même base de six images que dans [5]. Le contraste de Zeboudj a été calculé sur chaque résultat, il semble favoriser la sous-segmentation et ne paraît pas très exploitable. Malgré un réglage empirique des opérateurs, aucun algorithme ne s'est montré supérieur aux autres. Il est même impossible de classer les résultats de traitement d'une même image : ceci doit être fait par un spécialiste du domaine de l'image traitée, seul juge de la qualité du résultat en fonction du but précis qu'il cherche à atteindre.

3 Critères d'évaluation multi-échelles

Nous nous intéressons pour notre part au cas où l'on ne dispose pas de vérité-terrain et où l'on cherche à évaluer la qualité relative de différentes segmentations d'une même image.

Ce problème d'évaluation sans référence est extrêmement proche du problème de segmentation lui-même. En effet, de manière générale, le problème de segmentation (de partitionnement) d'image revient à savoir formuler une fonction de qualité sur les couples (image, partition) telle que la ou les partitions que l'on recherche (les résultats "acceptables" de segmentation) obtiennent la meilleure qualité. Ceci s'appuie sur un principe dit de "comparaison" qui, à notre connaissance, a été explicité pour la première fois par Koepfler, Lopez et Morel [13]: "We shall adopt a principle without which no discussion about segmentation can even start, and which we call comparison principle. It states that given two different segmentations of a datum, we are always able to decide which of them is considered as better than (or equivalent to) the other. Thus we assume the existence of some total ordering over all possible segmentations, and this can be simply acheived only if this ordering is reflected by some real functional E such that if $E(K_1) < E(K_2)$, then the segmentation K_1 has to be considered "better" than the segmentation K_2" . En d'autres termes, segmentation

et évaluation sans référence de résultats de segmentation passent tous deux par la définition d'une mesure de "qualité" Q sur les couples (image, partition). En segmentation, pour une image I donnée, on cherche à trouver la partition P qui maximise $Q(I, P)$ sur l'ensemble des partitions possibles du domaine de I. En évaluation, étant donnée une image I et un certain nombre de propositions de segmentation P_1, ..., P_k de I, on cherche i qui maximise $Q(I, P_i)$, ce qui permet de décréter que la segmentation P_i est la "meilleure". D'un point de vue théorique, le problème de segmentation et celui de l'évaluation sont donc similaires et reviennent au problème de savoir quantifier la qualité d'une segmentation. Toutefois pour obtenir un algorithme effectif de segmentation il faut savoir optimiser le critère de qualité et le choix des énergies de segmentation est profondément guidé par la capacité à optimiser efficacement l'énergie (de manière exacte ou approchée). Ce qui différencie donc en pratique les deux tâches "duales" de segmentation et d'évaluation est qu'en évaluation on pourra formuler des énergies bien plus complexes qu'en segmentation car on n'aura besoin que de calculer ces énergies sur quelques instances de segmentation et non à les optimiser sur l'espace des partitions possibles.

Si l'on s'intéresse à la mise au point de critères de qualité (i.e. d'énergies) pertinents, il semble alors judicieux de se poser la question sous l'angle de la segmentation : quelle(s) solution(s) trouverait-on si l'on optimisait le critère ? Ainsi, en segmentation, on cherche systématiquement à savoir si le problème d'optimisation est bien posé (au sens d'Hadamard) : possède-t-il une solution ? Si c'est le cas, cette solution est-elle unique ? Est-elle continue par rapport aux données ?

3.1 Défauts des critères existants

Examinons donc sous cet angle les critères d'évaluation sans référence décrits au paragraphe 2.2. Envisageons le cas d'une image uniforme (dont tous les pixels ont la même valeur) et demandons-nous quelle segmentation de cette image serait jugée la meilleure pour chacun des critères. On vérifie aisément que tous les critères sauf celui de Borsotti valent systématiquement zéro, quelle que soit la segmentation proposée de l'image uniforme ! Le critère de Borsotti, lui, attribue la meilleure note à la segmentation en une seule région, ce qui correspond au résultat attendu. Autrement dit, tous les critères sauf celui de Borsotti sont indécis dans le cas d'évaluation le plus trivial !

D'autre part, on sait que les formulations par optimisation d'un critère de contraste sont en général mal posées [23, 7, 10]. C'est pour cette raison que Zéboudj est obligé de définir une mesure par morceaux (équation 1): s'il s'était contenté de la formule $1 - I_i/E_i$ alors la sur-partition absolue aurait toujours le meilleur score (contraste interne nul). C'est pourquoi nous adopterons l'approche basée modèle.

3.2 Formulation énergétique générale et échelle

Nous posons le problème de segmentation d'image comme un problème de modélisation par morceaux d'une image (modélisation constante, polynomiale, gaussienne, lisse par morceaux, etc.) chaque région correspondant à un "morceau" du modèle. Une fois une classe de modèle sélectionnée (par exemple constant par morceaux), alors en vertu du principe de comparaison, la recherche du meilleur modèle peut toujours se formuler comme un problème d'optimisation : rechercher une partition R de l'image en régions et sur chaque région R_i de R le modèle M_{R_i} qui minimisent une certaine énergie totale $E(R)$. L'énergie doit évidemment prendre en compte la qualité de la modélisation en incorporant un terme $E_D(R)$ mesurant la distance entre le modèle et l'image. Toutefois, si l'on se contente d'une énergie d'adéquation du modèle aux données, la segmentation optimale est systématiquement la sur-segmentation absolue ou une partition proche. Par exemple, si l'on considère des modèles constants par morceaux, alors le modèle qui comporte une région par pixel, de valeur la valeur du pixel, est toujours solution exacte du problème, de distance nulle à l'image. Pour obtenir un résultat utile, il faut alors incorporer un terme énergétique $E_C(R)$, dit de "complexité", qui pénalise les segmentations trop fines, c'est-à-dire les modèles trop "complexes". Si l'on considère des modèles indépendants sur chaque région, on aboutit alors à des énergies qui prennent la forme générale [11]:

$$E(k, R) = \sum_{R_i \in R} E_D(R_i) + k \times E_C(R_i) \tag{5}$$

où k est un paramètre réel qui règle la contribution relative des deux termes énergétiques. Notons qu'outre contrôler la finesse de la solution, l'énergie de "complexité" E_C permet de contrôler la régularité géométrique de la solution, en privilégiant par exemple les régions aux frontières peu tortueuses. Dans un cadre probabiliste, l'énergie E_D peut s'envisager comme l'opposé de la log-vraisemblance des données sachant le modèle et l'énergie $k \cdot E_C$ comme l'opposé de la log-probabilité a priori du modèle. L'énergie totale est alors l'opposé de la log-probabilité a posteriori du modèle sachant les données et la minimiser revient à une recherche de maximum a posteriori. Une interprétation en tant que Lagrangien d'un problème d'optimisation sous contrainte, de type débit/distorsion, est également donnée dans [11].

Dans ce cadre, le choix d'une segmentation relève d'un compromis entre adéquation aux données et complexité du modèle et il n'y a pas intrinsèquement de meilleure solution : certaines applications peuvent avoir besoin d'un modèle précis, qui sera donc complexe, d'autres au contraire peuvent avoir besoin d'une description grossière, à grand traits, de l'image. Si l'énergie E_C est une fonction croissante avec la finesse de la partition, il est alors montré dans [9] que le paramètre k de l'équation (5) permet de contrôler la finesse de la solution, c'est-à-dire se comporte comme un *paramètre d'échelle* : comme nous l'avons vu, si $k = 0$ on trouve un modèle très morcelé qui s'adapte parfaitement à l'image, à l'inverse, pour k assez grand, l'image est modélisée par une seule

région. Suivant cette idée, il a été proposé dans [11, 9] de ne pas se contenter d'une seule partition minimisant (5) pour une valeur de k fixée a priori, mais de rechercher une famille de segmentations sous la forme d'une séquence $\{P_k\}_{k\in R^+}$ de partitions de finesse décroissante avec k. On montre que c'est équivalent à rechercher une hiérarchie indicée dont les ensembles représentent les régions appartenant aux minima de (5) et dont l'indice représente la plus petite échelle pour laquelle une région de la hiérarchie appartient à une solution du problème de minimisation. Un algorithme efficace pour trouver une séquence de solutions localement optimales, dans un sens précis, a été proposé dans [11, 9].

Nous inspirant de ce travail, nous nous intéressons ici à un problème complémentaire du problème de segmentation : celui de l'évaluation de la qualité de résultats de segmentation.

Nous reprenons pour celà l'expression énergétique à deux termes (5), attache aux données E_D et complexité E_C, et proposons de caractériser une segmentation $R = (R_1, \ldots, R_N)$ par la fonction $E(k, R)$ définie par :

$$k \longmapsto E(k, R) = \frac{1}{c}\sum_{i=1}^{N} E_D(R_i) + k \times \frac{1}{d}\sum_{i=1}^{N} E_C(R_i) \qquad (6)$$

où les coefficients c et d sont des coefficients de normalisation.

Il s'agit d'une fonction affine, croissante si $E_C = \sum_{i=1}^{N} E_C(R_i)$ est positive. Si l'on dispose de deux segmentations R et R' d'une même image, deux cas se présentent :

- soit $E_k(R) < E_k(R')$ pour tout k (ou inversement) auquel cas R est systématiquement meilleure que R' (ou inversement),
- soit il existe k_0 tel que $E_{k_0}(R) = E_{k_0}(R')$, auquel cas une des deux segmentations est meilleure aux petites échelles et l'autre meilleure aux grandes échelles.

Cette analyse se généralise au cas où l'on dispose d'un nombre arbitraire de segmentations. On montre facilement que l'ensemble des échelles k pour lesquelles une segmentation est meilleure que toutes les autres est un intervalle, éventuellement vide, qui représente la gamme d'échelles sur laquelle cette segmentation est la plus pertinente.

Notre approche permet donc d'ordonner les segmentations en fonction de l'échelle et d'évincer les segmentations qui ne sont pertinentes pour aucune échelle.

3.3 Différentes formes énergétiques

Différentes formes d'énergie d'attache aux données et d'énergie de "complexité" ont été étudiées dans [9].

3.3.1 Energie interne, d'attache aux données

Soit R_i une région comportant A_i pixels de valeurs $(X_1, X_2, ..., X_{A_i})$ et soit X_p^j la j-ème composante couleur du pixel p. Soit μ la moyenne des X_p et μ^j sa j-ème composante. Soit V la matrice de variance/covariance des X_p de terme général :

$$V(j,k) = \frac{1}{A_i} \sum_{p=1}^{A_i} \left(X_p^j - \mu^j\right) \left(X_p^k - \mu^k\right)$$

Pour l'attache aux données, l'énergie la plus basique est celle qui s'appuie sur un modèle constant par morceaux et évalue la fidélité modèle/image en norme L_2, aboutissant à

$$Q(R_i) = \sum_{p=1}^{A_i} \|X_p - \mu\|^2 = A_i \cdot Tr(V)$$

où $Tr(V)$ est la trace de V. Cette énergie a été initialement proposée par Mumford et Shah [20].

Une autre approche consiste à modéliser les valeurs de l'image au sein d'une région comme résultant d'un tirage i.i.d. selon une loi Gaussienne. En étendant aux dimensions supérieures l'approche par codage optimal développée en dimension 1 par Leclerc [15], on aboutit à une énergie proportionnelle à

$$G(R_i) = \ln(\det V) = \sum_{j=1}^{3} \ln(\lambda_j)$$

où les λ_j sont les valeurs propres de la matrice V.

Ce calcul n'est possible que si le déterminant est non nul, ce qui est faux dans de nombreux cas : régions de moins de 3 pixels, régions monochromes (où les 3 composantes couleur sont égales), régions dont une composante est fixe. On peut pallier ce problème en passant par le calcul des valeurs propres et en ramenant à 1 les valeurs propres inférieures à 1, pour un codage des couleurs entre 0 et 255 par canal.

Le calcul de la matrice de variance/covariance peut se faire dans n'importe quel espace couleur. Dans nos essais, nous avons conservé l'espace RVB d'origine. On notera toutefois que l'énergie G est invariante par rotation et homothétie du repère.

D'autres formes d'énergie interne très proches de la trace et du déterminant peuvent également être employées :

$$Q'(R_i) = A_i \cdot Tr(\sqrt{V}) = A_i \sum_{j=1}^{3} \sqrt{\lambda_j}$$

$$D(R_i) = A_i \cdot \det V = A_i \prod_{j=1}^{3} \lambda_j$$

$$D'(R_i) = A_i \cdot \det \sqrt{V} = A_i \prod_{j=1}^{3} \sqrt{\lambda_j}$$

Dans les expériences présentées ci-dessous, nous avons testé trois des formules ci-dessus : Q, G et D, et nous avons ramené à 1 les valeurs des déterminants ou des logarithmes des déterminants quand ils étaient inférieurs à 1. Pour des images codées sur $2n$ valeurs par composante, n^2 est un majorant des valeurs de variance et de covariance. Par conséquent le coefficient de normalisation c de l'équation (6) a été fixé à :

$$\begin{array}{l} c = n^2 \times 3 \times A_i \times 100 \text{ pour l'énergie } Q \\ c = n^6 \times 3 \times A_i \times 10000 \text{ pour l'énergie } D \\ c = A_i \text{ pour l'énergie } G \end{array} \quad (7)$$

3.3.2 Energie de complexité

En ce qui concerne l'énergie de "complexité", elle est calculée comme la somme sur toutes les régions d'une énergie qui peut être simplement constante, fournissant ainsi une énergie totale proportionnelle au nombre de régions de la segmentation.

Une autre forme simple d'énergie de complexité consiste à attribuer à chaque région une énergie proportionnelle à la longueur de sa frontière. L'énergie totale d'une partition correspond alors à la longueur totale des frontières de la partition, comme dans le modèle de Mumford et Shah.

D'autres énergies, calculées sur une approximation polygonale du contour ont été proposées dans [9] : nombre de côtés du polygone, concavité (somme des modules des angles) ou cohérence des directions des côtés adjacents. On peut également prendre en compte le gradient le long des frontières. L'énergie de complexité peut donc être simplement la longueur de la frontière :

$$L(R_i) = \sum_{s \in \delta R_i} 1$$

ou bien prendre en compte le gradient le long de la frontière :

$$LG(R_i) = \sum_{s \in \delta R_i} h(g(s))$$

où g représente le gradient et h est une fonction positive décroissante.

Le plus simple est de prendre $h(x) = \frac{1}{\|x\|}$, formule valide dès que le gradient est non nul, ce qui est en principe le cas sur les frontières.

Dans les expériences qui sont présentées ci-dessous, nous avons retenu l'énergie liée à la longueur de frontières. Quand l'image résultat présente les contours entre les régions, le calcul d'énergie s'effectue en comptant le nombre de pixels de contour. Pour les images résultats représentées sous la forme d'une image

de régions, sans contour, on calcule la longueur totale des frontières en comptabilisant le nombre de transitions verticales et horizontales entre pixels voisins appartenant à deux régions différentes. Le coefficient de normalisation d est égal au nombre de pixels de l'image. Le terme de "complexité" (avant multiplication par k) est donc compris entre 0 et 1.

4 Résultats expérimentaux

4.1 Comparaison de différentes énergies

Nous avons comparé les différentes énergies décrites ci-dessus sur plusieurs résultats de segmentation, obtenus par différents algorithmes. Les comparaisons sont d'abord effectuées sur l'image Maison (Fig.1), qui comporte des parties texturées et non texturées et pour laquelle la segmentation visuelle est relativement aisée et peut être consensuelle. Nous disposons de 6 résultats d'opérateurs de segmentation, les 5 premiers sont issus de [24]. Il s'agit de "Split and Merge" (SM), "Tominaga" (T), "Apprentissage compétitif" (A), "Croissance de régions" (C) et "Histogramme 2D" (H). Le sixième est obtenu par un algorithme de segmentation floue [21] (F). La figure 1 montre ces 6 résultats, chaque région étant représentée par la moyenne des couleurs des pixels qui la constituent. On distingue assez peu de différences entre les 6 images sauf pour Split and Merge sur laquelle un effet de bloc est visible. Ceci montre que les résultats présentés de cette façon sont souvent trompeurs ... La figure 2 les présente avec des niveaux aléatoires, sauf le ciel, le toit et la façade, qui sont représentés avec le même niveau pour toutes les segmentations. La méthode floue n'est pas très précise, les contours sont assez tortueux ; cependant le nombre de régions obtenues par cette segmentation est plus proche de la perception visuelle (voir Table 1) que pour les autres résultats. T, A, H comportent beaucoup de très petites régions, ces 3 résultats sont visuellement très proches. C contient aussi des très petites régions mais en nombre plus faible que les trois précédentes, et situées majoritairement sur les contours.

En première analyse la segmentation C semble fournir la segmentation la plus pertinente mais un examen plus attentif révèle la présence de nombreuses petites régions inutiles, si bien que F est aussi une bonne solution.

L'énergie que nous proposons est une somme pondérée de l'énergie interne d'attache aux données et d'une énergie de complexité égale à la longueur totale des contours (cf. Eq.(6)). Nous calculons d'abord séparément ces deux termes (énergie interne et énergie de complexité) et nous comparons les différentes formes d'énergie interne proposées. Le but est de vérifier si elles sont en correspondance avec notre perception visuelle.

Dans la figure 3, nous avons représenté sur un même graphique l'énergie de complexité et l'énergie interne, c'est-à-dire les couples (E_D, E_C) pour chacune des 6 segmentations. On observe que les méthodes Tominaga, Apprentissage compétitif, Histogramme 2D et Croissance de région fournissent des résultats très proches en terme d'énergie interne, quelle que soit la formule employée. La

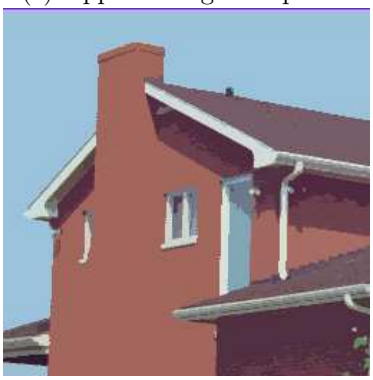
Image originale : Maison

(a) Split and merge

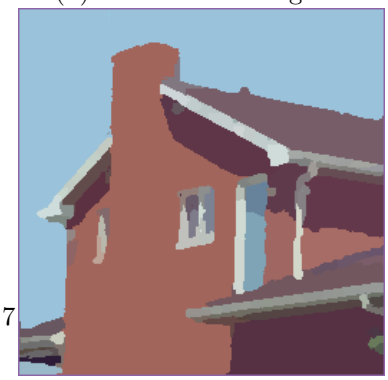
(b) Tominaga

(c) Apprentissage compétitif

(d) Croissance de régions

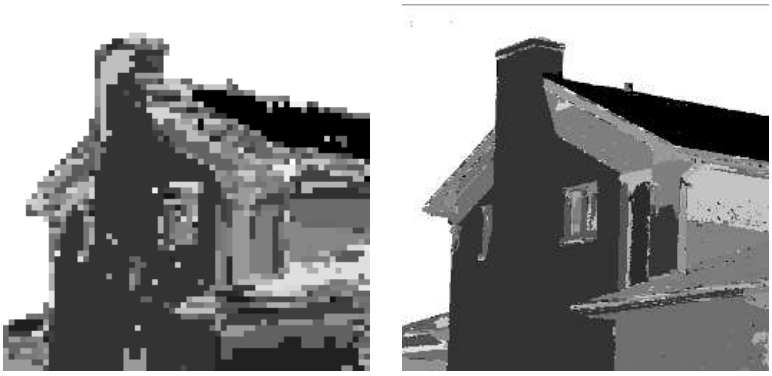


(e) Histogramme 2D



(f) Flou

Figure 1: 6 résultats de segmentation de l'image Maison. Chaque région est affectée de sa couleur moyenne



(a) Split and merge

(b) Tominaga

(c) Apprentissage compétitif

(d) Croissance de régions

(e) Histogramme 2D

(f) Flou

Figure 2: 6 résultats de segmentation de l'image Maison. Les régions sont représentées par des couleurs aléatoires

méthode floue diffère des autres pour les énergies internes Q et D, ainsi que pour l'énergie de complexité. La méthode floue se détache des autres pour les énergies Q et D, ainsi que par l'énergie de complexité. Split and Merge est bien séparée des autres résultats par l'énergie Q, et moins bien par les autres formes d'énergie interne.

Comme nous l'avons vu, le paramètre k de l'équation (5) est lié à la résolution ou au niveau de détail recherché. Les algorithmes qui donnent peu de régions ou des régions aux frontières régulières sont favorisés par l'énergie de complexité et défavorisés par l'énergie d'attache aux données. Nous comparons donc les segmentations en fonction du paramètre d'échelle k. Pour cela nous traçons (cf. Fig. 3d) pour chaque partition R la droite :

$$k \longmapsto E(k, R) = E_D(R) + k \cdot E_C(R)$$

Pour toutes les valeurs de k, la droite tracée pour C est toujours au-dessous de celles tracées pour A, T, SM et H. Donc ces quatre résultats sont pour toutes les échelles moins bonnes que C. La comparaison entre C et F varie selon le niveau de résolution. Pour une segmentation grossière, F est meilleur que C, pour une segmentation précise, C est meilleur. Finalement, parmi les 6 résultats de l'image Maison présentés, si on cherche une segmentation grossière, on optera pour le résultat F et pour une segmentation plus précise, on optera pour C.

Sur l'image Perroquet, nous disposons de 5 résultats de segmentation (cf. fig. 4). Les 4 premiers sont obtenus par classification des pixels, fuzzy C-means (FCM), une méthode neuro-floue (NF), Kohonen (KO) et k-means (KM) [19]. Le dernier est obtenu par segmentation floue (F) [21]. Visuellement les méthodes de classification sur-segmentent, alors que la méthode floue sous-segmente. De plus les résultats des méthodes FCM, KO et NF semblent très proches. Ce regroupement apparaît nettement sur les graphiques de la Fig. 5, surtout pour les énergies internes Q et G.

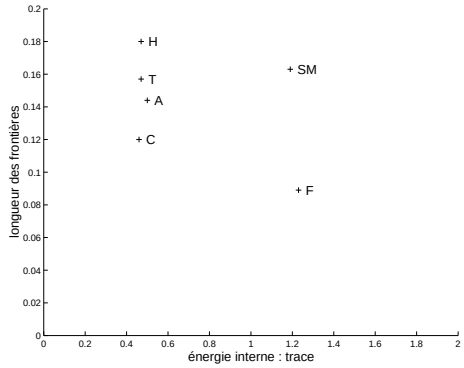
Les résultats obtenus sur de nombreux essais nous ont montré que l'énergie Q fournit les résultats correspondant le plus à notre appréciation visuelle. C'est donc elle que nous avons sélectionnée pour la suite de nos tests.

Le graphique 5d indique clairement la similitude des 3 résultats FCM, NF et KO à toutes les échelles. La méthode Kmeans n'est intéressante à aucune échelle. Pour une échelle grossière ($k > 10$) la méthode Floue fournit le meilleur résultat.

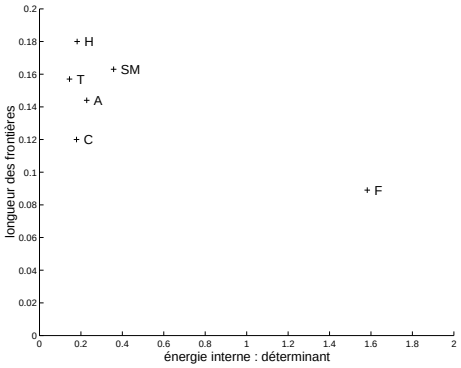
Comme nous l'avons dit plus haut, le paramètre k de l'Eq. (6) fixe l'échelle ou le degré de finesse recherché par la segmentation. Les algorithmes donnant peu de régions sont avantagés pour l'énergie de complexité et désavantagés pour l'énergie d'attache aux données.

L'énergie de complexité est très liée au nombre de régions, c'est pourquoi quand k est élevé, les méthodes qui comportent beaucoup de régions ont un score faible.

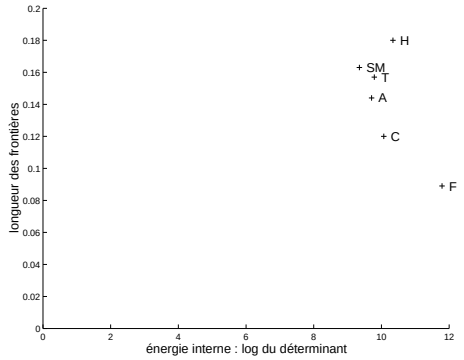
On peut fixer le facteur k en fonction du type de segmentation que l'on souhaite favoriser ; si on veut donner la préférence à une sur-segmentation (ou à une segmentation précise), on prendra k faible, inférieur à 10. Au contraire



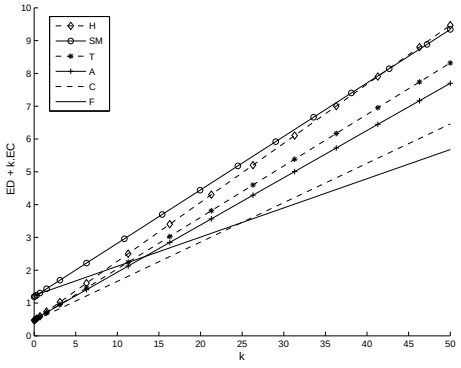
(a) Energie d'attache : Q



(b) Energie d'attache : D



(c) Energie d'attache : G



(d)$E(k, R)$ en fonction de k avec l'énergie Q

Figure 3: Energie de complexité par rapport à l'énergie d'attache aux données pour 6 résultats de segmentation de l'image Maison.

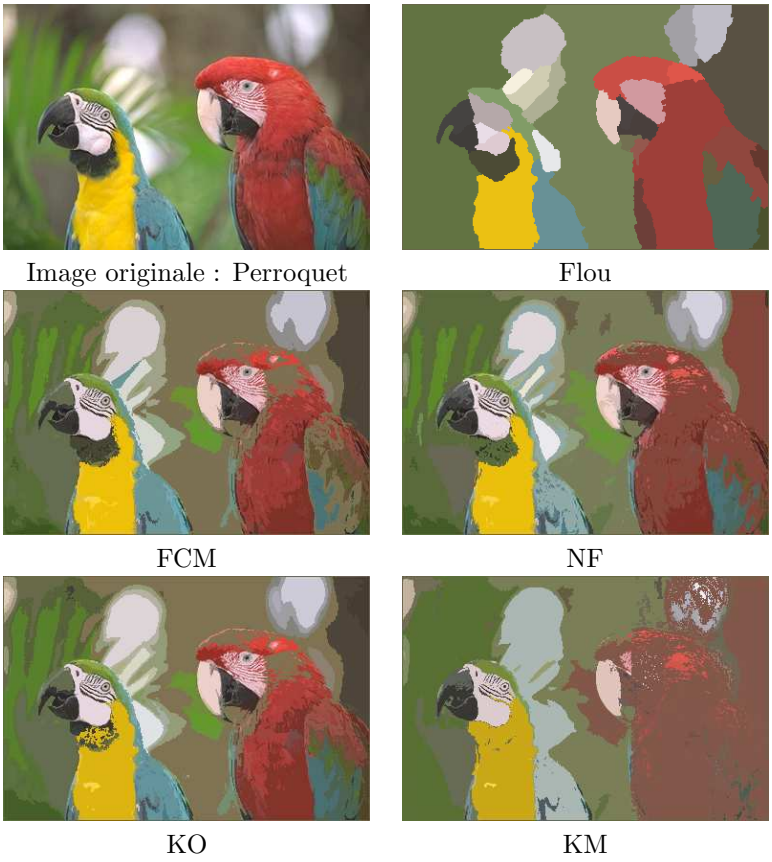
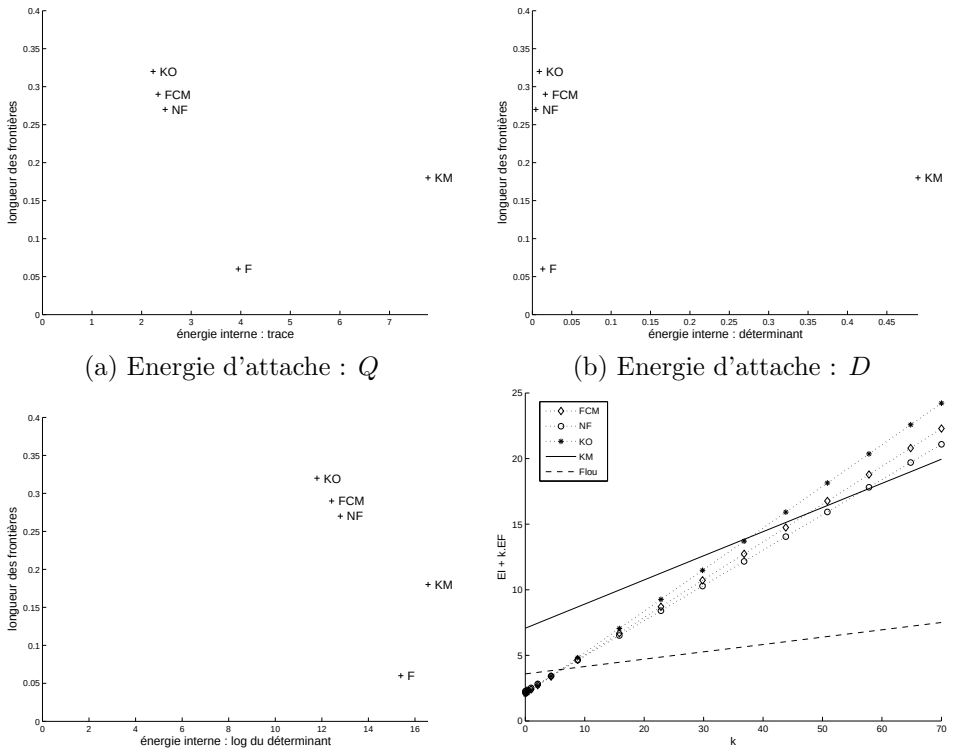


Figure 4: 5 résultats de segmentation de l'image Perroquet, chaque région est affectée de sa couleur moyenne

si on souhaite une segmentation grossière, avec peu de régions, on prendra une valeur de k supérieure à 10.



(a) Energie d'attache : Q

(b) Energie d'attache : D

(c) Energie d'attache : G

(d) $E(k, R)$ en fonction de k, énergie Q

Figure 5: Energie de complexité par rapport à l'énergie d'attache aux données pour chacun des 5 résultats de segmentation de l'image Perroquet.

4.2 Comparaison de différents critères d'évaluation

Dans un deuxième temps, nous avons comparé les principaux critères d'évaluation de la segmentation en régions : critère d'uniformité intra-région de Levine et Nazif, Liu et Yang, Borsotti et notre critère énergétique utilisant l'énergie Q avec différentes valeurs de k.

Une première comparaison a été effectuée sur les 6 résultats de segmentation de l'image Maison. Les tables 1 et 2 présentent les valeurs des différents critères pour chaque segmentation, avec en gras la méthode de segmentation qui fournit le meilleur score pour chaque critère.

	SM	T	A	C	H	F
nombre de régions	776	1375	1057	654	1384	139
Levine-Nazif	116	78	65	49	70	**31**
Liu-Yang	3.2	0.40	0.37	**0.25**	0.39	0.47
Borsotti	0.4	29	8	1.1	24	**0.1**
Energie $k = 10$	2.82	2.04	1.94	**1.66**	2.28	2.12
Energie $k = 100$	17.5	16.2	14.9	12.5	18.5	**10.2**

Table 1 : valeurs des critères pour chacun des 6 résultats de segmentation de l'image Maison

	FCM	NF	KO	KM	Flou
nombre de régions	2952	2748	3398	1708	34
Levine et Nazif	95	49	57	18	**7**
Liu et Yang	1.18	1.06	1.29	1.36	**0.18**
Borsotti	198	171	360	38	**0.11**
Energie $k = 10$	5.01	4.94	5.20	8.91	**4.14**
Energie $k = 100$	30.96	29.11	33.70	25.42	**9.21**

Table 2 : Valeurs des critères pour chacun des 5 résultats de segmentation de l'image Perroquet

Tous les critères (sauf Borsotti) donnent des valeurs proches pour les images visuellement proches, T, A, C et H pour la Maison obtiennent des scores de même ordre, et, pour l'image Perroquet, on retrouve les trois groupes mentionnés plus haut, les classifications floues d'une part, les segmentations floues d'autre part et Kmeans dans un troisième groupe. L'échelle choisie pour le critère énergétique, favorise la sous-segmentation et place bien en tête les résultats de segmentation floue. Le critère de Borsotti est extrêmement sensible aux petites régions d'un ou deux pixels, à cause du deuxième terme de l'Eq. 4.

Une seconde comparaison des différents critères d'évaluation a été effectuée sur des images dont on possède des segmentations "manuelles". Elles sont issues de la base de Berkeley (cf. §2.1.2)

Pour l'image "Tulipe" (cf. fig. 6 et table 3), les 3 critères Levine-Nazif, Liu-Yang et Borsotti préfèrent la segmentation (a), qui est une caricature en 4 régions de l'image. Ces 3 critères privilégient donc les segmentations grossières. Comme le montre la figure 7, pour notre critère multi-échelles, la segmentation (e) est jugée la meilleure pour les petites échelles; aux échelles moyennes c'est la segmentation (c), et aux grandes échelles c'est la segmentation (a) qui arrive en tête. Les segmentations (b) et (d), elles, ne sont jamais jugées pertinentes. On constate en effet visuellement qu'elles sont hétérogènes en terme de niveau de détail : elles sont globalement grossières mais très détaillées en certains lieux précis.

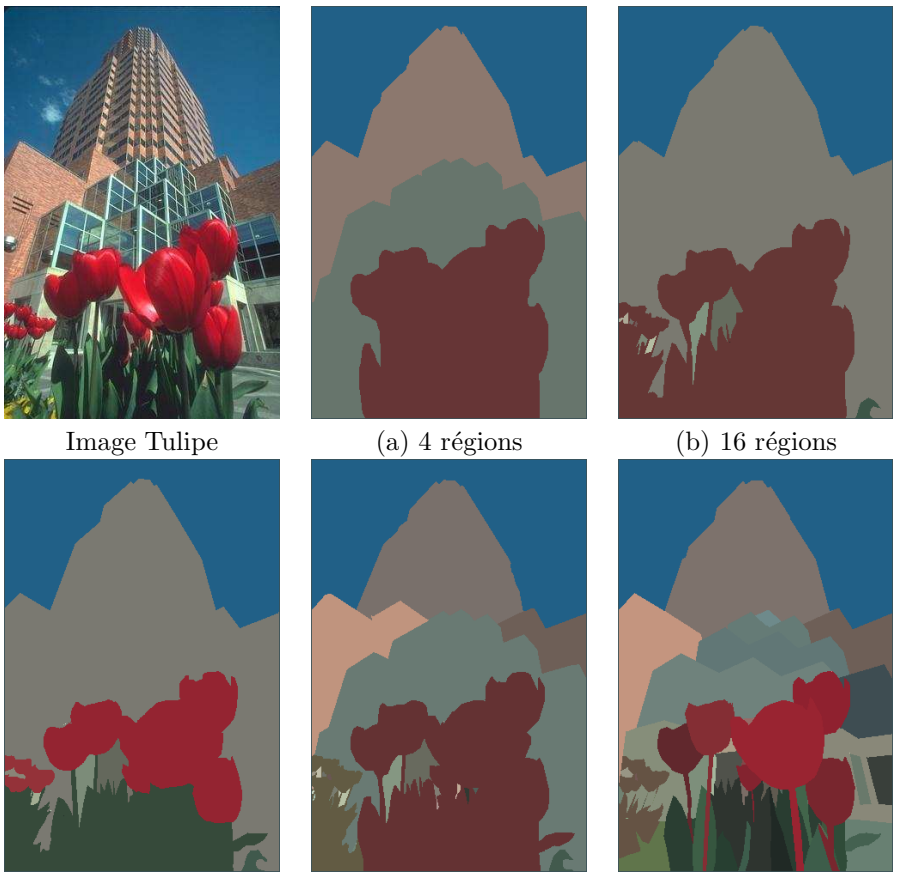


Image Tulipe (a) 4 régions (b) 16 régions

(c) 20 régions (d) 41 régions (e) 79 régions

Figure 6: 5 segmentations manuelles de l'image "Tulipe" de la base de Berkeley.

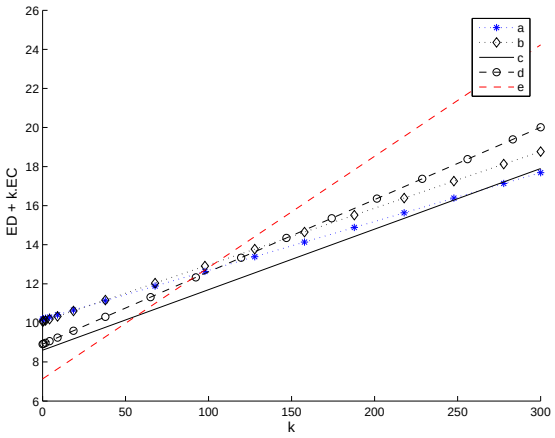


Figure 7: E_k en fonction de k (énergie d'attache Q) pour les 5 segmentations manuelles de la figure 6 (image "Tulipe").

	a	b	c	d	e
nombre de régions	4	16	20	41	79
Levine-Nazif	**2.76**	4.56	6.31	8.78	14.9
Liu-Yang	**0.06**	0.10	0.11	0.20	0.46
Borsotti	**0.1**	0.17	0.17	0.26	0.32
Energie $k = 10$	10.45	10.36	8.91	9.29	**7.70**
Energie $k = 100$	**12.7**	13	11.73	12.65	12.8

Table 3 : Comparaison des critères pour les 5 segmentations manuelles de l'image Tulipe.

Nous avons extrait de la base de Berkeley (cf. §2.1.2) l'image présentée Fig. 8 et les 5 segmentations manuelles de cette image. Ces segmentations sont très grossières, les détails ne sont extraits dans aucune des images, et pour certaines, seul le contenu sémantique a été recherché et non une séparation objective selon la couleur par exemple. C'est ainsi que pour les segmentations (c) et (e), le personnage forme une seule région. L'image (d) est la plus précise, en comparaison les 4 autres sous-segmentent. Mais elles sous-segmentent le fond pour les images (a) et (b) et le personnage pour les images (c) et (e). On constate d'autres différences entre ces segmentations manuelles. La table 4 fournit les résultats des différents critères. L'image (d) qui est la plus précise, est considérée comme la meilleure par les critères de Levine et Nazif et notre critère énergétique pour $k = 10$. On peut conclure avec notre critère que si on souhaite beacoup de détail, il faut choisir la segmentation (d) et si on souhaite

une segmentation moins précise, il faut choisir le résultat (a) ou (b) qui sont très proches visuellement.

	a	b	c	d	e
nombre de régions	33	69	48	72	43
Levine et Nazif	9.5	10.9	**6.25**	15.68	9.02
Liu et Yang	**0.16**	0.26	0.20	0.25	0.18
Borsotti	**0.14**	0.20	0.18	0.15	0.18
Energie $k = 10$	5.26	5.26	9.46	**3.87**	6.19

Table 4 : Valeurs des critères pour chacun des 5 résultats de segmentation manuelle de l'image de la Fig. 8

5 Conclusion

L'évaluation des algorithmes de traitement d'images et notamment ceux de segmentation constitue un problème d'importance, tant pour le choix d'un algorithme et de son paramétrage pour un utilisateur que pour la comparaison avec l'existant d'un nouvel algorithme par un chercheur. Le nombre important de critères quantitatifs témoigne d'un besoin de toute la communauté en traitement d'images.

Nous avons tenté de recenser ces critères pour les algorithmes de segmentation en régions, que l'on ait ou non une vérité-terrain.

Nous nous sommes ensuite focalisés sur les méthodes de segmentation en régions d'images couleur pour lesquelles nous ne possédons pas de vérité-terrain, ce qui nous semble être le problème le plus général et le plus difficile à traiter. De nombreux critères d'évaluation existent dans la littérature, tous prenant en compte l'écart à la moyenne des couleurs des pixels constituant la région et pour certains (Liu-Yang et Borsotti) cherchant à modéliser la complexité de l'image segmentée par le nombre de régions. Il nous a paru nécessaire de prendre en compte la résolution attendue par l'utilisateur et qui dépend de son but. Il est en effet illusoire de vouloir comparer deux segmentations qui l'une extrait grossièrement les régions, dans le but par exemple de vérifier la présence d'un objet ou de compter des occurrences d'objets et l'autre s'efforce d'extraire avec précision tous les détails de tous les objets présents dans l'image. Il est donc capital d'évaluer un résultat de segmentation en fonction d'un but, dont un des éléments quantifiables est le niveau de résolution souhaité.

C'est pourquoi nous avons proposé des critères d'évaluation liés à la résolution et qui prennent en compte à la fois la complexité de la segmentation et la fidélité des régions extraites à l'image de départ. Le premier aspect est mesuré par la longueur des contours, qui permet de quantifier à la fois le nombre de régions (lié à la résolution) et la régularité des contours. Le deuxième aspect est l'attache aux données quantifiée pour un modèle gaussien par plusieurs expressions assez proches les unes des autres. Après comparaison, il semble que le plus efficace parmi ces critères soit le critère de Mumford et Shah.

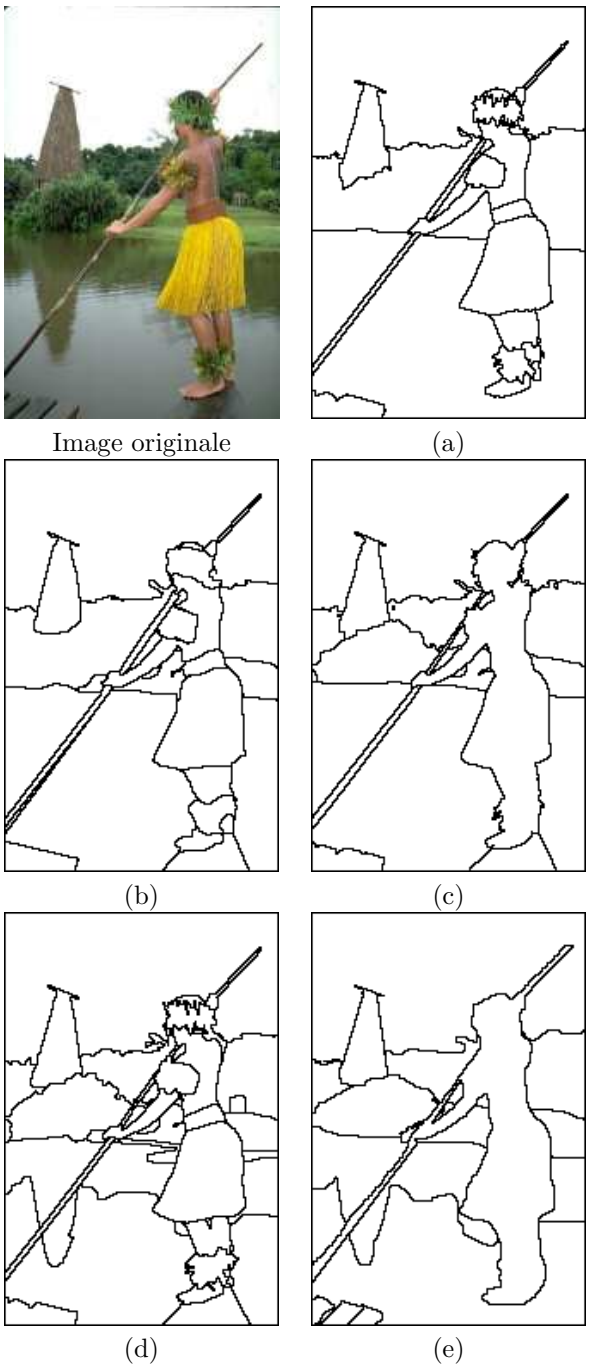


Figure 8: Segmentations manuelles sur une image de la base de Berkeley

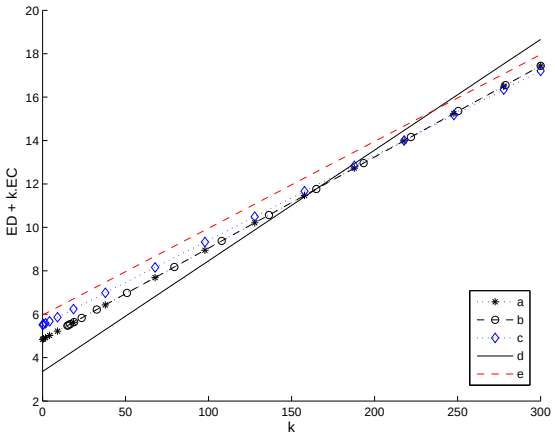


Figure 9: $E(k, R)$ en fonction de k (énergie d'attache Q) pour les 5 segmentations manuelles de la figure 8.

Ces nouveaux critères sont de plus très simples à calculer sur tout type d'images, monochromes ou multispectrales, avec ou sans contours entre les régions.

Remerciements : Nous remercions vivement Ludovic Macaire et Fella Hachouf pour leurs résultats de segmentation respectivement sur Maison et Perroquet, ainsi que David Picard et Jérôme Dantan pour leur programmes d'évaluation.

References

[1] M. Borsotti, P. Campadelli, and R. Schettini. Quantitative evaluation of color image segmentation results. *Pattern Recognition Letters*, 19:741–747, 1998.

[2] J. S. Cardoso and L. Corto-Real. Toward a generic evaluation of image segmentation. *IEEE trans. on Image Processing*, 14(11), 2005.

[3] S. Chabrier, C. Rosenberger, H. Laurent, B. Emile, and P. March. Evaluating the segmentation result of a gray-level image. In *12th EUSIPCO*, Vienne, Austria, september 2004.

[4] V. Chalana and Y. Kim. A methodology for evaluation of boundary detection algorithms on medical images. *IEEE Trans. on Medical Imaging*, 16(5):642–652, 1997.

[5] J.P. Cocquerez and S. Philipp editors. *Analyse d'images: filtrage et segmentation*. Masson, Paris, 1995.

[6] D. Coquin, P. Bolon, and B. Ionescu. Dissimilarity measures in color spaces. In *16th ICPR*, volume 1, 2002.

[7] P. Felzenszwalb and D. Huttenlocher. Efficiently computing a good segmentation. In *Conf. Computer Vision and Pattern Recognition*, Santa Barbara, CA, USA, 1998.

[8] L. Guigues. Image segmentation comparison using hierarchical model for n-m region matching. In *Proc. of 2nd IAPR TC-15 Workshop on Graph-based Representations in Pattern Recognition, Haindorf, Autria*, 1999.

[9] L. Guigues. *Modèles multi-échelles pour la segmentation d'images*. PhD thesis, Université de Cergy-Pontoise, 2003.

[10] L. Guigues, H. Le Men, and J.-P. Cocquerez. The hierarchy of the cocoons of a graph and its application to image segmentation. *Pattern Recognition Letters*, 24(3):1024–1066, 2003.

[11] L. Guigues, H. Le Men, and J.-P. Cocquerez. Scale-sets image analysis. In *Proc. of IEEE Int. Conf. on Image Processing (ICIP'03), Barcelona, Spain*, September 2003.

[12] F. Huet and S. Philipp. Fusion of images interpreted by a new fuzzy classifier. *Pattern Analysis and Applications*, 1:230–247, 1998.

[13] G. Koepfler, Lopez, and J.-M. Morel. A multiscale algorithm for image segmentation by variational method. *SIAM journal on numerical analysis*, 31(1):282–299, 1994.

[14] H. Laurent, S. Chabrier, C. Rosenberger, B. Emile, and P. Marché. Etude comparative de critères d'évaluation de la segmentation. In *19th GRETSI*, Paris, France, june 2003.

[15] Y. Leclerc. Constructing simple stable descriptions for image partitioning. *Int. J. of Computer Vision*, 3(1):73–102, 1989.

[16] M.D. Levine and A.M. Nazif. Dynamic measurement of computer generated image segmentations. *IEEE Trans. on PAMI*, 7(25):155–164, 1985.

[17] J. Liu and Y.-H. Yang. Multiresolution color image segmentation. *IEEE Trans. on PAMI*, 16(7):689–700, 1994.

[18] D. R. Martin. *An empirical approach to grouping and segmentation*. PhD thesis, University of California, Berkeley, USA, 2002.

[19] N. Mezhoud, F. Hachouf, and M. Batouche. Segmentation dimages couleurs par une méthode neuro-floue. In *Taima03*, pages 170–176, Hammamet, Tunisie, 2003.

[20] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and asociated variational problems. *Comm. Pure Appl. Math.*, 42:577–685, 1989.

[21] S. Philipp-Foliguet, M. B. Vieira, and M. Sanfourche. Fuzzy segmentation of color images and indexing of fuzzy regions. In *First Europ. conf. on Colour in Graphics, Imaging and Vision*, pages 507–512, Poitiers, France, 2002.

[22] P.K. Sahoo, S. Soltani, and A.K.C. Wong. A survey of thresholding technique. *Comput. Vision Graphics Image Process*, 4:233–260, 1988.

[23] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE trans. on PAMI*, 22(8):888–905, August 2000.

[24] A. Trémeau, C. Fernandez-Maloigne, and P. Bonton. *Image numérique couleur*. Dunod, Paris, 2004.

[25] L. Vinet. *Segmentation et Mise en Correspondance de Régions de Paires d'Images Stéréoscopiques*. PhD thesis, Université Paris IX - Dauphine, July 1991.

[26] D. L. Wilson, A. J. Baddeley, and R. A. Owens. A new metric for grey-scale image comparison. *Int. J. of Computer Vision*, 24:5–17, 1997.

[27] W. A. Yasnoff, W. Galbraith, and J. W. Bacus. Error measures for objective assessment of scene segmentation algorithms. *AQC*, 1:107–121, 1979.

[28] S. Yu and J. Shi. Segmentation given partial grouping constraints. *IEEE trans. on Image Processing*, pages 173–183, 2004.

[29] Y. J. Zhang. A survey on evaluation mehods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.