# évaluation de segmentations d'images : mesures de similarité avec une référence partielle

Paméla Daum, Jean-Luc Buessler, Jean-Philippe Urban

# Évaluation de segmentations d'images : mesures de similarité avec une référence partielle

# A supervised evaluation method for image segmentation: discrepancy measures with a partial reference

Paméla DAUM Jean-Luc BUESSLER Jean-Philippe URBAN

MIPS, Université de Haute-Alsace
4, rue des frères Lumière, 68093 Mulhouse Cedex France
{pamela.daum, jean-luc.buessler, jean-philippe.urban}@uha.fr

## Résumé

*La validation d'une technique de segmentation d'images nécessite généralement de disposer d'exemples de référence. La création de ces segmentations de référence, réalisée manuellement sur des images représentant de nombreux objets, est longue, fastidieuse et rarement précise. La référence présente toujours des simplifications introduites par l'expert.*
*Cet article évalue la possibilité d'utiliser une référence délibérément simplifiée, qui ne représente qu'un nombre réduit d'objets les plus intéressants pour caractériser l'algorithme de segmentation. L'étude présentée porte principalement sur les critères de comparaison de l'image segmentée avec une référence partielle.*
*L'intérêt de l'approche est illustré dans le contexte de l'analyse d'images de microscopie électronique.*

## Mots Clef

Segmentation d'images, évaluation supervisée, référence partielle.

## Abstract

*Reference segmentations are generally needed to validate an image segmentation algorithm. Constructing these references manually on images containing many objects is a tedious, time-consuming task that often lacks in precision. The reference is always somehow simplified by the expert. This paper introduces a reference that is intentionally simplified. It represents a reduced number of the most interesting objects to characterize the segmentation algorithm. The paper presents adapted supervised criteria for an automatic segmentation comparison with a partial reference. The interest of the method is illustrated for the analysis of electron-microscope images.*

## Keywords

Image segmentation, supervised evaluation, partial reference.

## 1 Introduction

La segmentation est souvent une première étape pour l'analyse automatique des images. Elle prépare les traitements ultérieurs en groupant les pixels en régions [6]. La pertinence de cette partition est importante et doit tenir compte aussi bien de la nature de l'image que des objectifs de l'analyse.

La pertinence d'une segmentation peut être évaluée en s'appuyant sur diverses approches, la plupart recensées par Zhang [15]. Les méthodes non-supervisées [14, 4] évaluent une segmentation par rapport aux caractéristiques des segments : les critères utilisés sont des critères d'homogénéité, de forme, de taille.

Les méthodes supervisées [3, 12] évaluent des segmentations d'images dont les propriétés sont complexes ou inconnues. Elles comparent plusieurs segmentations automatiques à une segmentation de référence et classent les segmentations suivant leur ressemblance à la référence. Dans le cas où aucune vérité terrain n'est disponible, la référence peut être une segmentation manuelle.

Pour valider ou améliorer un algorithme de segmentation, il est indispensable de comparer un nombre important d'images. Une image comprenant un grand nombre d'objets est longue et difficile à segmenter manuellement. La segmentation manuelle d'une grande quantité d'images est une tâche fastidieuse et l'expert est alors tenté de ne segmenter que les contours les plus significatifs.

La complexité des images de microscopie électronique que nous traitons nous a conduits à simplifier le travail demandé à l'expert. Nous lui proposons de créer des segmentations de référence partielles en ne marquant que quelques objets importants dans chaque image. Les objets peuvent être choisis parce qu'ils sont caractéristiques de l'application, mais aussi pour éprouver l'algorithme de segmentation. Par exemple, pour vérifier que des objets faiblement contrastés sont correctement identifiés.

Cet article a pour objectif de proposer une adaptation des

critères de comparaison de segmentations à une référence partielle. La méthode d'évaluation proposée combine des critères déjà existants avec une définition originale des mesures de sur-segmentation et de sous-segmentation.
La deuxième partie de ce papier situe le contexte de notre étude. La troisième partie porte sur l'état de l'art de l'évaluation supervisée de segmentations et des critères utilisés. La quatrième partie présente la définition d'une évaluation à l'aide d'une référence partielle et propose une discussion sur l'adaptation des critères décrits dans l'état de l'art. À partir de cette discussion, de nouveaux critères sont définis. La cinquième partie évalue, sur des images réelles, les caractéristiques des critères. Enfin les conclusions et perspectives sont énoncées.

## 2 Contexte de l'étude

Cette étude s'inscrit dans le projet européen HT3DEM [2, 7] dédié à l'automatisation complète d'une chaîne de cristallisation 2D de protéines membranaires. La cristallisation 2D est une technique importante pour l'analyse structurelle de protéines qui ne peuvent s'organiser en cristaux 3D. L'analyse d'images issues d'un microscope électronique à transmission permet de déterminer si une membrane est cristalline. Cependant, le nombre d'expériences requises pour déterminer de bonnes conditions de cristallisation est important. C'est pourquoi, les images sont acquises et analysées automatiquement.
La phase la plus importante dans le processus d'analyse d'images est la segmentation. En effet, c'est de celle-ci que dépend la caractérisation des membranes. Il est nécessaire d'évaluer cette segmentation par rapport à son objectif : détecter toutes les membranes.

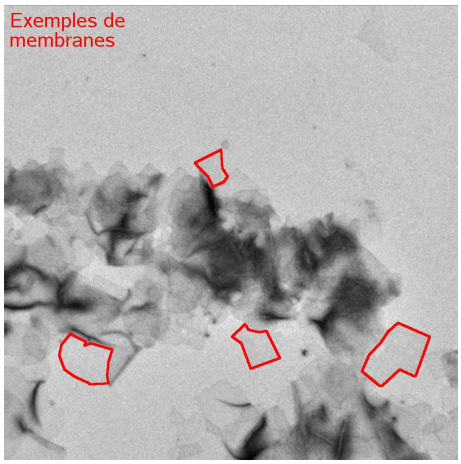



FIG. *1 : Exemple d'image de microscopie électronique*

Les images analysées contiennent des objets dont les caractéristiques sont complexes (cf. figure 1) : les formes et la taille sont diverses, les membranes peuvent être superposées, il n'y a pas de textures significatives ni d'uniformité à l'intérieur des régions et le contraste inter-région peut être faible. Pour évaluer la segmentation automatique de ces images (p.e. figure 4e), on la compare à une segmentation manuelle.
Or, le nombre d'objets est élevé et il est quasiment impossible pour un expert de segmenter de manière précise toutes les membranes. En général, seules les membranes significatives sont segmentées par l'expert. Par ailleurs, dans l'évaluation d'un algorithme, les contours bien contrastés ne sont pas discriminants et les différences induites par des contours mal identifiés par l'expert sont peu intéressantes. Ces remarques nous ont conduits à proposer une évaluation automatique de segmentations en utilisant une référence composée de quelques objets les plus significatifs et difficiles à segmenter.

## 3 État de l'art de l'évaluation supervisée

L'évaluation supervisée est une méthode empirique qui consiste à comparer une segmentation automatique $\hat{S} = \bigcup_{j=1}^{l} \hat{R}_j$ à une segmentation de référence $S = \bigcup_{i=1}^{k} R_i$. $S$ et $\hat{S}$ sont des partitions de l'image : tous les pixels sont indexés et appartiennent à une région $R_i$ de $S$ et à une région $\hat{R}_j$ de $\hat{S}$.
Les techniques d'évaluation utilisent, pour la plupart [9], des mesures de différences basées sur une matrice de confusion :

$$T_{ij} = \mathrm{card}(R_i \cap \hat{R}_j), \tag{1}$$

où, $R_i$ est la $i^e$ région dans la segmentation de référence et $\hat{R}_j$ est la $j^e$ région dans la segmentation automatique. La matrice, aussi appelée table d'appariement, recense tous les couples de régions et leur nombre de pixels de recouvrement.
Comparer une segmentation par rapport à une référence consiste à vérifier que chaque région de la référence est correctement représentée dans la segmentation. Puisque les indices sont arbitraires, il faut apparier les régions les plus semblables avant d'évaluer leur similarité. Le tableau d'appariement permet de trouver ces paires de régions grâce à leur aire de recouvrement.
La mesure de VINET (utilisée dans [5]) se fonde sur l'appariement itératif des régions dont le recouvrement est maximal :

$$C_1 = \max\{T_{i_1 j_1}\}, \tag{2}$$

$$\vdots$$

$$C_x = \max\{T_{i_x j_x}\}, \tag{3}$$

$$\vdots$$

$$C_m = \max\{T_{i_m j_m}\}, \tag{4}$$

avec $i_1 = 1 \dots k$, $j_1 = 1 \dots l$ et $i_x \neq i_1 \dots i_{x-1}$, $j_x \neq j_1 \dots j_{x-1}$ et $m = \min\{k, l\}$. La mesure de dissimilarité de VINET est définie par :

$$d_V = \frac{\mathrm{card}(S) - \sum_{x=1}^{m} C_x}{\mathrm{card}(S)}. \tag{5}$$

C'est une distance simple qui, cependant, ne donne aucune indication sur le degrée de sur- ou de sous-segmentation.

HOOVER [10] énonce des critères de classification de chacune des régions de référence. Un premier critère détermine les régions bien segmentées. On note $\theta_{cd}(i,j,t)$ une instance de bonne détection d'une paire de régions, en fixant un seuil de recouvrement minimum $0{,}5 < t \leq 1$.

$$T_{ij} \geq (t \times \mathrm{card}(R_i)) \text{ et } T_{ij} \geq (t \times \mathrm{card}(\hat{R}_j)). \quad (6)$$

Soit $E_j$ l'ensemble des régions $j_1 \dots j_n$ avec $2 \leq n \leq l$. $\theta_{os}(i, E_j, t)$ est une instance de sur-segmentation de la région $R_i$ représentée par plusieurs régions $\hat{R}_j$, si :

$$\sum_{x \in E_j} T_{ix} \geq (t \times \mathrm{card}(R_i)), \quad (7)$$

$$\text{avec } E_j = \{j | T_{ij} \geq (t \times \mathrm{card}(\hat{R}_j))\}. \quad (8)$$

Soit $E_i$ l'ensemble des régions $i_1 \dots i_m$ avec $2 \leq m \leq k$. $\theta_{us}(E_i, j, t)$ est une instance de sous-segmentation de plusieurs régions $R_i$ représentées par $\hat{R}_j$, si :

$$\sum_{x \in E_i} T_{xj} \geq (t \times \mathrm{card}(\hat{R}_j)), \quad (9)$$

$$\text{avec } E_i = \{i | T_{ij} \geq (t \times \mathrm{card}(R_i))\}. \quad (10)$$

Les deux autres critères de classification de HOOVER sont les suivants : $\theta_m(i,t)$ est une instance de non-détection si la région $R_i$ n'est comprise dans aucune instance $\theta_{cd}$, $\theta_{os}$, $\theta_{us}$. $\theta_n(j,t)$ est une instance de bruit si la région $\hat{R}_j$ n'est comprise dans aucune instance de détection, sur- ou sous-segmentation.

La classification de HOOVER caractérise ainsi chaque région segmentées. HOOVER est le premier à notre connaissance utilisant un seuil de proportion de recouvrement $0{,}5 < t \leq 1$ qui permet de donner une certaine tolérance à la segmentation.

ORTIZ [12] s'inspire de HOOVER en utilisant le seuil de recouvrement et introduit des mesures prenant en compte la taille des régions. Il énonce ainsi trois mesures globales :

$$CG = \sum_{i=1}^{k} \sum_{j=1}^{l} SRa(i,j,t) \times T_{ij}\,; \quad (11)$$

$$US = \sum_{j=1}^{l} (1 - SRb(j,t)) \times \mathrm{card}(\hat{R}_j)\,; \quad (12)$$

$$OS = \sum_{i=1}^{k} (1 - SO(i,t)) \times \mathrm{card}(R_i)\,, \quad (13)$$

avec

$$SRa(i,j,t) = \begin{cases} 1 & \text{si } T_{ij} \geq (t \times \mathrm{card}(\hat{R}_j) \\ 0 & \text{sinon}\,; \end{cases} \quad (14)$$

$$SRb(j,t) = \begin{cases} 1 & \text{si } \max\limits_{x=1,\cdots,k} \{T_{xj}\} \geq (t \times \mathrm{card}(\hat{R}_j)) \\ 0 & \text{sinon}\,; \end{cases} \quad (15)$$

$$SO(i,t) = \begin{cases} 1 & \text{si } \max\limits_{x=1,\cdots,l} \{T_{ix}\} \geq (t \times \mathrm{card}(R_i)) \\ 0 & \text{sinon.} \end{cases} \quad (16)$$

Les trois mesures $CG$, $US$ et $OS$ indiquent respectivement le taux de pixels bien détectés, le taux de sous-segmentation et le taux de sur-segmentation. Pour les normaliser, elles sont divisées par le nombre total de pixels de l'image originale.
La prochaine section discute la pertinence des critères de HOOVER et des mesures d'ORTIZ lorsqu'une segmentation doit être comparée à une référence partielle.

# 4 Comparaison avec une référence partielle

## 4.1 La référence partielle

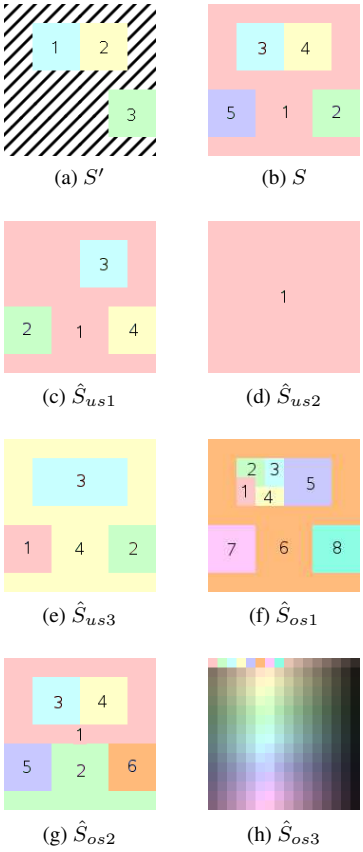
Soit une segmentation de référence partielle $S'$ comprenant $k$ objets d'intérêt, un objet étant défini par une seule région. On note $I = 1, \dots, k$ l'ensemble des indices des régions choisies, $S' = \bigcup_{i \in I} R_i$ l'ensemble de ces objets et $\mathrm{card}(R_I) = \sum_{i \in I} \mathrm{card}(R_i)$. $S'$ a les propriétés suivantes :
- $\forall i \in I, R_i \neq \emptyset$ ;
- $\forall a, b \in I^2, a \neq b, R_a \cap R_b = \emptyset$.

Seuls les objets d'intérêt sont indexés dans la référence. $S'$ n'est donc pas une partition de l'image, les pixels hors des régions d'intérêt ne sont pas pris en considération.
On définit dans $\hat{S}$ l'ensemble $J$ des indices des régions recouvrant $S'$ et $\mathrm{card}(\hat{R}_J) = \sum_{j \in J} \mathrm{card}(\hat{R}_j)$. La table d'appariement $T'$ devient :

$$T'_{ij} = \mathrm{card}(R_i \cap \hat{R}_j), \text{avec } i = 1 \in I \text{ et } j \in J. \quad (17)$$

## 4.2 Pertinence des critères

Pour adapter les critères d'ORTIZ aux domaines de définition respectifs des $R_i$ et $\hat{R}_j$, les mesures globales sont normalisées en divisant $CG$ et $OS$ par $\mathrm{card}(R_I)$ et $US$ par $\mathrm{card}(\hat{R}_J)$.
Pour vérifier la pertinence des critères, on utilise une image de synthèse comportant un nombre de régions connus. Leur taille et leur degré de sur-segmentation et sous-segmentation sont également connus. La figure 2 représente des segmentations de l'image de synthèse $S$ comportant 5 objets dont quatre sont à l'intérieur du cinquième. Ce sont des segmentations étiquetées et les régions n'ont pas forcément les mêmes indices d'une segmentation à l'autre.




(a) $S'$ (b) $S$ (c) $\hat{S}_{us1}$ (d) $\hat{S}_{us2}$ (e) $\hat{S}_{us3}$ (f) $\hat{S}_{os1}$ (g) $\hat{S}_{os2}$ (h) $\hat{S}_{os3}$

FIG. *2 : Segmentations d'une image de synthèse $16 \times 16$ : les résultats sont indexés et les couleurs représentent l'indice de chaque région*

Une référence partielle $S'$ est choisie (figure 2a) pour l'évaluation, les pixels n'intervenant pas sont hachurés.

Avec une comparaison visuelle, voici ce que l'on peut dire des segmentations :

- $S$ : segmentation idéale ;
- $\hat{S}_{us1}$ : $R_1$ n'est pas détectée ;
- $\hat{S}_{us2}$ : aucun objet de référence n'est détecté ;
- $\hat{S}_{us3}$ : $R_1$ et $R_2$ sont fusionnées par $\hat{R}_3$ ;
- $\hat{S}_{os1}$ : $R_1$ est divisée par $\hat{R}_1$, $\hat{R}_2$, $\hat{R}_3$ et $\hat{R}_4$ ;
- $\hat{S}_{os2}$ : $\hat{R}_1$ et $\hat{R}_2$ n'affectent aucun objet de référence ;
- $\hat{S}_{os3}$ : chaque pixel de l'image est un segment.

Les résultats de la comparaison à l'aide des critères d'ORTIZ sont présentés dans le tableau 1. Une segmentation idéale $S$ comparée à $S'$ (ligne 1 du tableau 1) est correctement évaluée : la corrélation $CG$ est maximale, les mesures de sous-segmentation $US$ et de sur-segmentation $OS$ sont à 0.

L'analyse du tableau 1 nous montre que seule la mesure $OS$ est proportionnelle au taux de sur-segmentation, les mesures $CG$ et $US$ sont mal adaptées à ce contexte d'utilisation.

Lorsqu'il y a une ou plusieurs régions non-détectées ($\hat{S}_{us1}$ et $\hat{S}_{us2}$), $US$ mesure la sous-segmentation mais elle est proportionnelle à la taille de $\hat{R}_j$ et non à la taille de la région de référence. $US$ ne fait pas la distinction entre sous-segmentation ($\hat{S}_{us3}$) et non-détection.

$CG$ ne pénalise pas la sur-segmentation comme le montre les résultats des trois dernières lignes du tableau 1.

Par contre, $OS$ est bien adaptée puisque lorsqu'il y a sur-segmentation, la mesure est proportionnelle à la taille de la région de référence. On peut également noter que lorsqu'une région non choisie est sur-segmentée comme dans $\hat{S}_{os2}$, $OS$ est nulle : la mesure ne se fait que sur les régions choisies comme référence.

De part les défauts de $US$ et $CG$, les critères d'ORTIZ ne peuvent pas être utilisés avec une référence partielle.

L'évaluation qualitative proposée par HOOVER reste, par contre, pertinente et classe correctement l'ensemble des régions testées. Avec $t = 0{,}8$ on peut notamment vérifier :

- $S' \backslash \hat{S}_{us1}$ : $R_1$ est une instance de non-détection $\theta_m(1, t)$ ;
- $S' \setminus \hat{S}_{us3}$ : $R_1$ et $R_2$ participent à l'instance de sous-segmentation $\theta_{us}(E_i, 3, t)$ avec $E_i = \{1, 2\}$ ;
- $S' \setminus \hat{S}_{os3}$ : aucune région ne participe à une instance de bonne détection.

Les critères de HOOVER prennent en compte l'ensemble des régions qui se superposent partiellement avec un objet de référence. Ils n'imposent pas que les segmentations soient des partitions complètes de l'image et sont donc parfaitement adaptés à nos objectifs.

Cependant, ces critères classent chacune des régions sans fournir de mesures globales pour l'évaluation et la comparaison de segmentations.

| | $t = 0{,}8$ | | |
|---|---|---|---|
| **Comparaisons** | $CG$ | $US$ | $OS$ |
| $S' \setminus S$ | 1 | 0 | 0 |
| $S' \setminus \hat{S}_{us1}$ | 0,67 | 0,78 | 0 |
| $S' \setminus \hat{S}_{us2}$ | 0 | 1 | 0 |
| $S' \setminus \hat{S}_{us3}$ | 0,33 | 0,67 | 0 |
| $S' \setminus \hat{S}_{os1}$ | 1 | 0 | 0,33 |
| $S' \setminus \hat{S}_{os2}$ | 1 | 0 | 0 |
| $S' \setminus \hat{S}_{os3}$ | 1 | 0 | 1 |

TAB. *1 : Mesures obtenues des critères d'ORTIZ pour la comparaison de $S'$ avec les autres segmentations de la figure 2*

## 4.3 Critères proposés

Pour établir ses mesures, ORTIZ classe chacune des régions selon 3 critères puis réalise la somme, pondérée par leur taille, de toutes les régions respectant un critère. Nous reprenons ce principe en y apportant deux modifications. D'une part, la classification des régions repose sur les critères de HOOVER, mieux adaptés. D'autre part, la contribution d'une région à un critère n'est plus considérée comme binaire (tout ou rien), mais caractérisée plus finement par

une proportion, par exemple un taux de fractionnement ou un taux de similarité.
Cette approche permet d'améliorer l'évaluation d'une segmentation à partir d'un nombre réduit de régions.
$\theta_{cd}(i,j,t)$, $\theta_{os}(i,E_j,t)$, $\theta_{us}(E_i,j,t)$ et $\theta_m(i,t)$ peuvent être aisément adaptés aux domaines de définition respectifs des $R_i$ et $\hat{R}_j$.
La mesure associée à une bonne détection indique le taux de similarité. Si la région $R_i$ participe à une instance $\theta_{cd}(i,j_1,t)$ alors :

$$RC(i,t) = \min\{\frac{T'_{ij_1}}{\text{card}(R_i)}, \frac{T'_{ij_1}}{\text{card}(\hat{R}_{j_1})}\}. \quad (18)$$

Notons que $RC(i,t) > t$ et tend vers 1 lorsque $R_i$ et $\hat{R}_{j_1}$ se superposent exactement. $RC(i,t) = 0$ si $R_i$ n'est pas bien détectée.
La deuxième mesure indique le taux de sur-segmentation ou fragmentation. Si la région $R_i$ participe à une instance $\theta_{os}(i,E_j,t)$ alors :

$$RF(i,t) = 1 - \frac{\sum_{j\in E_j}(T'_{ij} \times (T'_{ij}-1))}{\text{card}(R_i) \times (\text{card}(R_i)-1)}. \quad (19)$$

Cette mesure, inspirée par l'indice de diversité de SIMPSON [13], correspond à la probabilité que deux pixels différents de $R_i$ appartiennent à des régions $\hat{R}_j$ différentes.
La troisième mesure indique le taux de sous-segmentation ou fusion. Si la région $R_i$ fait partie de $E_i$ dans une instance $\theta_{us}(E_i,j_1,t)$ alors :

$$RA(i,t) = 1 - \frac{\sum_{x\in E_i}(T'_{xj_1} \times (T'_{xj_1}-1))}{\text{card}(R_{E_i}) \times (\text{card}(R_{E_i})-1)}. \quad (20)$$

De la même manière que $RF(i,t)$, la mesure correspond à la probabilité que deux pixels différents de $R_{E_i}$ appartiennent à des régions $R_i$ différentes.
L'évaluation de la non-détection reste binaire. Si $R_i$ participe à une instance $\theta_m(i,t)$, alors :

$$RM(i) = 1. \quad (21)$$

On peut énoncer les mesures globales $RC$, $RF$, $RA$ et $RM$ pour l'ensemble des régions choisies dans une image ou dans un ensemble d'images :

$$\overline{RC(t)} = \frac{1}{\text{card}(R_I)} \times \sum_{i\in I}(RC(i,t) \times \text{card}(R_i)); \quad (22)$$

$$\overline{RF(t)} = \frac{1}{\text{card}(R_I)} \times \sum_{i\in I}(RF(i,t) \times \text{card}(R_i)); \quad (23)$$

$$\overline{RA(t)} = \frac{1}{\text{card}(R_I)} \times \sum_{i\in I}(RA(i,t) \times \text{card}(R_i)); \quad (24)$$

$$\overline{RM(t)} = \frac{1}{\text{card}(R_I)} \times \sum_{i\in I}(RM(i,t) \times \text{card}(R_i)). \quad (25)$$

La section suivante montre des exemples de l'utilisation de ces nouveaux critères sur des images réelles.

# 5 Évaluation des régions d'intérêt

Dans cette partie, l'utilisation d'une référence partielle et des critères d'évaluation proposés est illustrée et discutée d'une part sur une image extraite de la base de Berkeley [11], d'autre part sur l'image de microscopie électronique présentée au début de l'article.

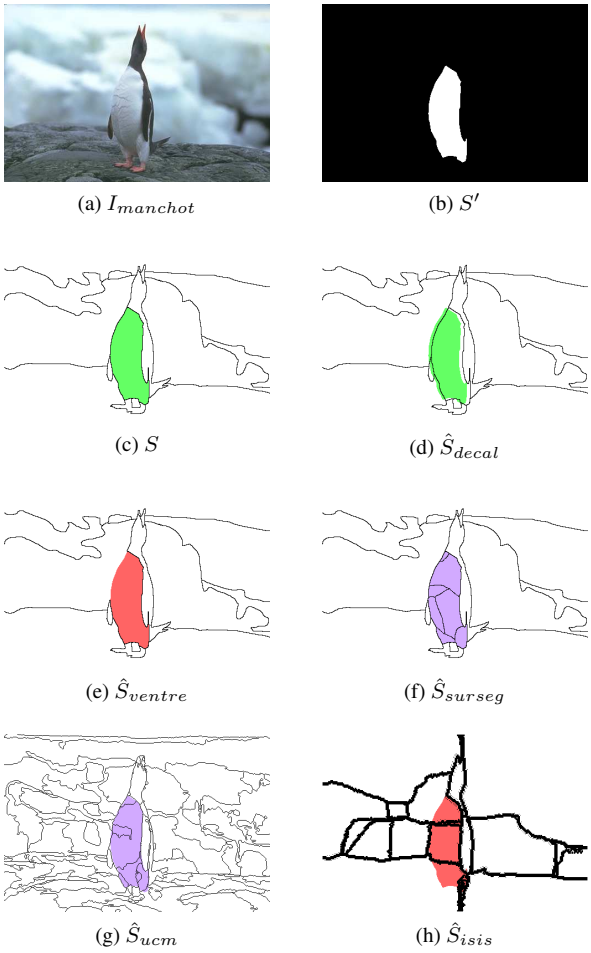
## 5.1 Évaluation d'une région d'intérêt

La première image $I_{manchot}$, présentée figure 3, a été choisie parce qu'elle présente une région difficile à segmenter : il s'agit d'un manchot dont le ventre se confond avec les icebergs à l'arrière plan de l'image. La référence est une segmentation manuelle mise à disposition sur le site de Berkeley [1]. La référence partielle retenue pour cette première étude comporte une région unique correspondant à la zone ventrale claire. Pour les comparaisons, le seuil de recouvrement est de $t = 0{,}8$.
Les lignes 2 et 3 de la figure 3 présentent des résultats manuellement modifiés afin de montrer certaines caractéristiques de l'évaluation.
Les images sur la dernière ligne de la figure 3 montrent le résultat de la comparaison avec d'autres segmentations disponibles sur le site de Berkeley et réalisés avec les algorithmes « Ultrametric Contour Maps » ($\hat{S}_{ucm}$) et « Segmentation Induced by Scale Invariance » ($\hat{S}_{isis}$). Il faut rappeler que les critères analysés évaluent des segmentations de l'image définies comme une indexation des régions et ne s'appliquent qu'à des contours fermés.
Pour apprécier visuellement les résultats des critères, un code couleur a été appliqué aux résultats de segmentation. Ce code est défini dans le tableau 2.

| Couleurs | Critère | Description du critère |
|---|---|---|
| Vert | $RC$ | Bonne corrélation |
| Violet | $RF$ | Région fractionnée |
| Jaune | $RA$ | Région agglomérée |
| Rouge | $RM$ | Non détection |

TAB. *2 : Code des couleurs utilisées dans la représentation des résultats de comparaison*

Lorsque la référence partielle est comparée à la segmentation manuelle, il n'y a pas d'erreur. Lorsque le manchot est un peu décalé, la similarité est de $RC(1,t) = 0{,}87$. Cette valeur diminue lorsque le décalage augmente : en dessous du seuil $t$, la région est considérée comme une instance de non-détection.
Lorsque le ventre n'est pas détecté ($\hat{S}_{ventre}$) ou lorsqu'une seule partie est segmentée ($\hat{S}_{isis}$), la région est considérée comme non-détectée.
Lorsque le ventre est sur-segmenté ($\hat{S}_{surseg}$ et $\hat{S}_{ucm}$), la région est considérée comme fragmentée et les mesures sont $RF_{surseg}(1,t) = 0{,}77$ et $RF_{ucm}(1,t) = 0{,}62$. Bien



(a) $I_{manchot}$ (b) $S'$ (c) $S$ (d) $\hat{S}_{decal}$ (e) $\hat{S}_{ventre}$ (f) $\hat{S}_{surseg}$ (g) $\hat{S}_{ucm}$ (h) $\hat{S}_{isis}$

FIG. *3 : Image originale du manchot, image de référence et segmentations associées*

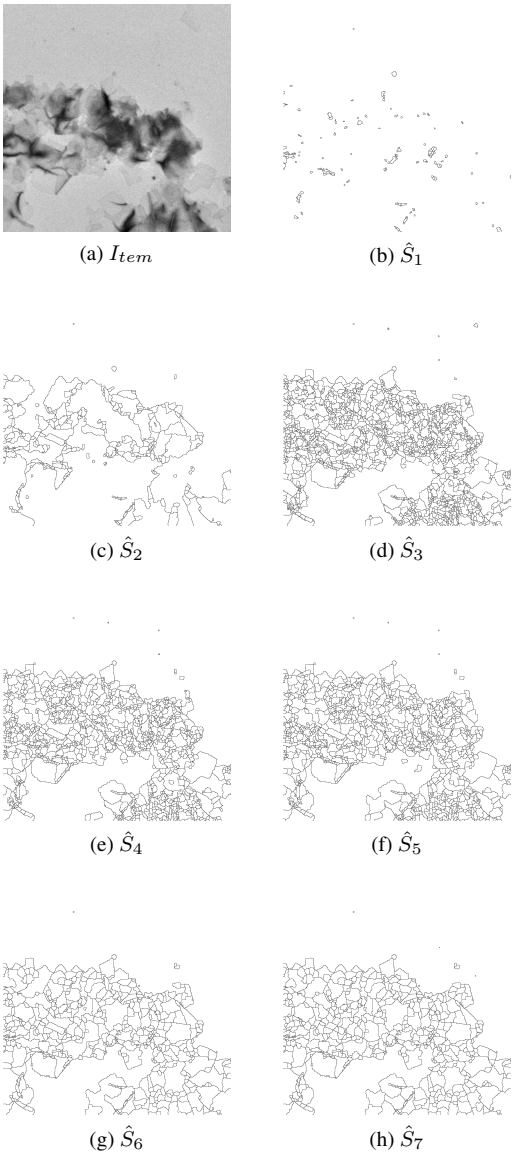
que dans les deux segmentations, la zone ventrale apparaisse fractionnée en 7 régions, la diversité mesurée dans $RF_{ucm}(1,t)$ est plus faible puisqu'il y a une grande région (et 6 autres plus petites), tandis que dans $\hat{S}_{surseg}$, deux régions sont grandes. La mesure $RF$ évalue la sur-segmentation en tenant compte du nombre et de la taille des régions assimilées à la région de référence.

Cet exemple ne présente pas de cas d'agglomération, puisque l'objet de référence est unique.

Grâce aux critères $RC$, $RF$, $RA$, et $RM$, on peut avoir le taux de similarité, de sur-segmentation et de sous-segmentation, d'une ou plusieurs régions de référence.

## 5.2 Évaluation de plusieurs régions d'intérêt

La deuxième image a été acquise par un microscope électronique et représente des membranes potentiellement cristallines. On peut voir que certaines sont difficilement distingables par rapport au fond : ce sont ces membranes que l'on va retenir pour l'évaluation. La segmentation partielle de référence a été faite manuellement et les résultats de segmentation à tester sont issus d'une segmentation automatique multi-résolution [8]. Le seuil utilisé pour la comparaison des segmentations est $t = 0{,}75$.



(a) $I_{tem}$ (b) $\hat{S}_1$ (c) $\hat{S}_2$ (d) $\hat{S}_3$ (e) $\hat{S}_4$ (f) $\hat{S}_5$ (g) $\hat{S}_6$ (h) $\hat{S}_7$

FIG. *4 : Image originale de microscopie électronique et segmentations associées*

La figure 4 présente l'image originale issue du microscope électronique ainsi que 7 résultats de segmentation. Ces résultats sont classés suivant la préférence visuelle qu'on leur accorde : $\hat{S}_1$ est la partition la moins bonne, et $\hat{S}_7$ est la meilleure. Ce classement est subjectif et peut être discuté. Il prend en compte l'objectif qui est de segmenter un maximum de membranes. La meilleure segmentation est donc celle qui aura détecté le plus de membranes avec le moins de sur-segmentation possible.

Ainsi le classement se justifie de la sorte : $\hat{S}_1$ et $\hat{S}_2$ comportent peu ou pas de membranes. $\hat{S}_3$, $\hat{S}_4$ et $\hat{S}_5$ présentent davantage de contours, ce qui peut être un signe de sur-segmentation et il manque certaines membranes. $\hat{S}_6$ et $\hat{S}_7$ présentent moins de contours que les précédentes segmen-

tations mais ont détecté plus de membranes.

La segmentation manuelle précise et exhaustive de l'ensemble de l'image est une tâche trop fastidieuse pour être demandée à un expert. La référence, représentée figure 5a, est composée de 8 régions.

Les résultats numériques de l'évaluation des 7 segmentations par rapport à cette référence sont présentés dans le tableau 3. Les deux segmentations $\hat{S}_1$ et $\hat{S}_2$ sont inadéquates puisque le taux de non-détection $\overline{RM}$ est égal ou proche de 1.

| | $S'$ **comparée avec** | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\hat{S}_1$ | $\hat{S}_2$ | $\hat{S}_3$ | $\hat{S}_4$ | $\hat{S}_5$ | $\hat{S}_6$ | $\hat{S}_7$ |
| $\overline{RC}$ | 0 | 0,04 | 0,61 | 0,66 | 0,51 | 0,43 | 0,46 |
| $\overline{RF}$ | 0 | 0 | 0,14 | 0,02 | 0,05 | 0,23 | 0,32 |
| $\overline{RA}$ | 0 | 0 | 0 | 0 | 0,13 | 0 | 0 |
| $\overline{RM}$ | 1 | 0,96 | 0,29 | 0,29 | 0,14 | 0,25 | 0 |

TAB. *3 : Mesures globales pour la comparaison des segmentations avec $S'$*

La seule segmentation à présenter un cas de sous-segmentation est $\hat{S}_5$.

Ce tableau 3 nous permet de faire trois classements suivant les taux de similitude, de sur-segmentation et de non-détection, du taux le moins bon au meilleur :

– $\overline{RC}$ : $\hat{S}_1$, $\hat{S}_2$, $\hat{S}_6$, $\hat{S}_7$, $\hat{S}_5$, $\hat{S}_3$, $\hat{S}_4$ ;
– $\overline{RF}$ : $\hat{S}_7$, $\hat{S}_6$, $\hat{S}_3$, $\hat{S}_5$, $\hat{S}_4$, $\hat{S}_1 = \hat{S}_2$ ;
– $\overline{RM}$ : $\hat{S}_1$, $\hat{S}_2$, $\hat{S}_3 = \hat{S}_4$, $\hat{S}_6$, $\hat{S}_5$, $\hat{S}_7$.

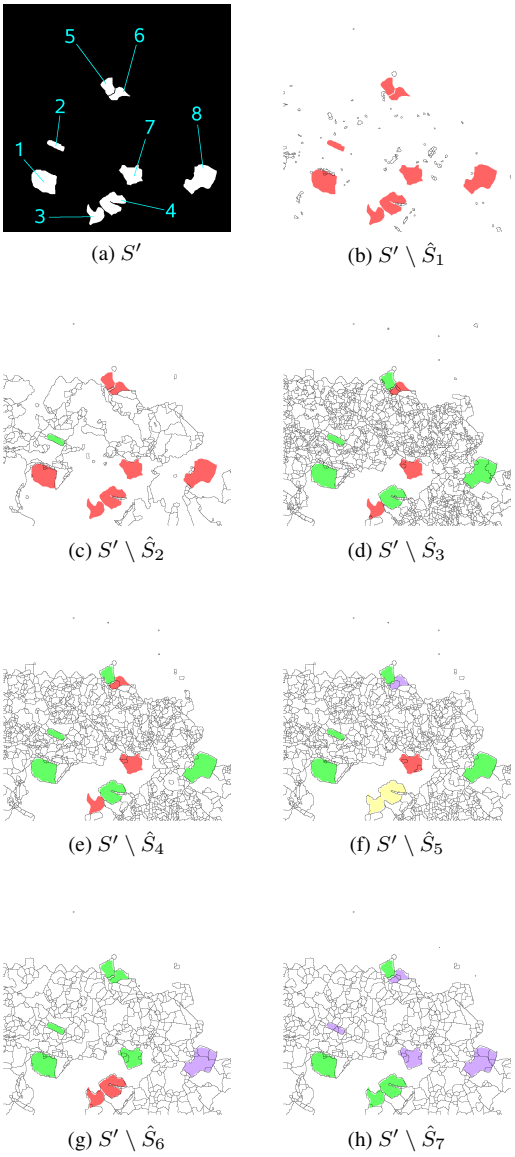
Ces classements indiquent que $\hat{S}_5$ et $\hat{S}_7$ ont détecté un maximum de membranes mais que $\hat{S}_7$ présente le plus de sur-segmentation. On pourrait conclure que $\hat{S}_5$ est la meilleure segmentation.

Cependant, le contexte de notre application nous conduit à moins pénaliser la sur-segmentation puisque les régions sont quand même détectées et seront évaluées par des post-traitements. La figure 5 représente les résultats de la comparaison avec le même code couleur que celui énoncé dans le tableau 2. Visuellement, $\hat{S}_7$ semble être la meilleure segmentation. Dans cette analyse, la mesure de sur-segmentation devient moins importante que la mesure de sous-segmentation. Par conséquent, d'après les valeurs du tableau 3, $\hat{S}_7$ devient la meilleure segmentation.

Cette méthode de comparaison implique différents critères. Ceci laisse le choix à l'utilisateur d'établir des règles de sélection de segmentation par rapport à l'application.

# 6 Conclusions et perspectives

La segmentation est une phase importante dans l'analyse d'images. Son évaluation est indispensable afin de vérifier sa pertinence par rapport aux objectifs d'une application. Pour des images dont les caractéristiques des objets sont inconnues ou complexes, l'évaluation se fait grâce à la comparaison avec une référence. Dans la plupart des cas, la référence est une segmentation manuelle. Celle-ci est compliquée à réaliser si l'image comporte beaucoup d'objets.



(a) $S'$ (b) $S' \setminus \hat{S}_1$ (c) $S' \setminus \hat{S}_2$ (d) $S' \setminus \hat{S}_3$ (e) $S' \setminus \hat{S}_4$ (f) $S' \setminus \hat{S}_5$ (g) $S' \setminus \hat{S}_6$ (h) $S' \setminus \hat{S}_7$

FIG. *5 : Référence $S'$ de l'image $I_{tem}$ et comparaisons des diverses segmentations*

Cet article propose une simplification de la référence en n'évaluant que certains objets de l'image. Pour utiliser une référence partielle, nous définissons des mesures de comparaison originales en nous inspirant des travaux d'ORTIZ et de HOOVER. Pour chaque objet de la référence, les régions les mieux appariées sont identifiées : en fonction de leur nombre, l'objet est classé comme bien identifié, fractionné, aggloméré ou non détecté dans la segmentation. Des taux de similarité, de fragmentation ou d'agglomération apportent une indication plus précise sur la reconnaissance des objets. Des mesures globales synthétisent l'information sur l'ensemble de l'image.

L'utilisation de ces nouvelles mesures est illustrée sur des images réelles. Un premier exemple discute la classifica-

tion et les mesures en étudiant une région unique, la zone ventrale du manchot. Un exemple de l'évaluation de segmentations d'une image de microscopie électronique est ensuite proposé. Huit régions membranaires composent la référence.
Ces outils de mesure permettent de comparer des segmentations en utilisant une référence partielle. Ils caractérisent globalement et individuellement les objets bien détectés, sur-segmentés ou sous-segmentés.
Une évaluation plus systématique de cette approche est en cours, notamment dans le contexte du traitement des images en microscopie électronique.

## Remerciements

Ce travail est soutenu par EU 6th framework (HT3DEM, LSHG-CT-2005-018811). Les images de microscopie ont été acquises par l'équipe du Biozentrum, Bâle.

# Références

[1] The berkeley segmentation dataset and benchmark. http ://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/.

[2] High throughput - three dimensional electron microscopy. http ://www.ht3dem.org/.

[3] S. Chabrier. Évaluation de résultats de segmentation d'images. In *JJC LVR'04*, 2004.

[4] S. Chabrier, B. Emile, C. Rosenberger, and H. Laurent. Unsupervised Performance Evaluation of Image Segmentation. *Applied Signal Processing*, 1 :1–12, 2006.

[5] S. Chabrier, C. Rosenberger, and B. Emile. Évaluation de la performance de la segmentation d'images par fusion de critères. In *ORASIS'05*, 2005.

[6] J.-P. Cocquerez and S. Philipp. *Analyse d'images : filtrage et segmentation*. Masson, 1997.

[7] N. Coudray, F. Beck, J.-L. Buessler, A. Korinek, H. Rémigy, H. Kihl, J. M. Plitzko, and J.-P. Urban. Automation for the Acquisition and Integrated Evaluation of Electron Micrographs from 2D-Crystallization of Proteins. In the Microscopy Society of America and the Microbeam Analysis Society, editors, *Microscopy and Microanalysis Conference*, volume 13, pages 422–423, Ft. Lauderdale, Florida, 2007.

[8] N. Coudray, J.-L. Buessler, H. Kihl, and J.-P. Urban. Multi-scale and First Derivative Analysis for Edge Detection in TEM Images. In Springer Berlin Heidelberg, editor, *ICIAR'07*, volume 4 633 of *Lecture Notes in Computer Science*, pages 1 005–1 016. Montreal, Canada, 2007.

[9] J. Da Rugna and H. Konik. Étude comparative de méthodes de segmentation dans une approche orientée indexation. In *RFIA'04*, pages 13–20, Toulouse, 2004.

[10] A. Hoover, G. Jean-Baptiste, X. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher. An Experimental Comparison of Range Image Segmentation Algorithms. *Pattern Analysis and Machine Intelligence*, 18(7) :673–689, 1996.

[11] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In IEEE Computer Society, editor, *ICCV'01*, volume 2, pages 416–425, Vancouver, Canada, 2001.

[12] A. Ortiz and G. Oliver. On the use of the overlapping area matrix for image segmentation evaluation : A survey and new performance measures. *Pattern Recognition Letters*, 27(16) :1 916–1 926, 2006.

[13] E.H. Simpson. Measurement of Diversity. *Nature*, 163 :688, 1998.

[14] H. Zhang, J. E. Fritts, and S. A. Goldman. Image Segmentation Evaluation : A survey of Unsupervised Methods. *Computer Vision and Image Understanding*, 110(2) :260–280, 2008.

[15] Y. J. Zhang. A Survey on Evaluation Methods for Image Segmentation. *Pattern Recognition*, 29(8) :1 335–1 346, 1996.