Évaluation d'un résultat d'interprétation d'images

Baptiste Hemery* ** ***, Hélène Laurent****,
Bruno Emile**** et Christophe Rosenberger* ** ***

*Université de Caen Basse-Normandie, UMR 6072 GREYC, F14032 Caen, France
**ENSICAEN, UMR 6072 GREYC, F14032 Caen, France
{baptiste.hemery, christophe.rosenberger}@ensicaen.fr
http://www.ecole.ensicaen.fr/~hemery/
http://www.ecole.ensicaen.fr/~rosenber/
***CNRS, UMR 6072 GREYC, F14032 Caen, France

****Laboratoire PRISME, ENSI de Bourges - Université d'Orléans
88 boulevard Lahitolle, 18020 Bourges - France
helene.laurent@ensi-bourges.fr, bruno.emile@univ-orleans.fr

Résumé. Les algorithmes de traitement d'images regroupent un ensemble de méthodes qui vont traiter l'image depuis son acquisition jusqu'à l'extraction de l'information utile pour une application donnée. Parmi ceux-ci, les algorithmes d'interprétation ont pour but de détecter, localiser et/ou reconnaître un ou plusieurs objets dans une image. Le problème traité réside dans l'évaluation de résultats d'interprétation d'une image ou une vidéo lorsque l'on dispose de la vérité terrain associée. Les enjeux sont multiples comme la comparaison d'algorithmes, l'évaluation d'algorithmes en cours de développement ou leur paramétrage optimal. Cet article présente la méthode d'évaluation de la qualité d'un résultat d'interprétation d'image que nous avons développée. Cette méthode permet de prendre en compte la qualité de la localisation, de la reconnaissance ainsi que de la détection des objets. Paramétrable, cette méthode peut être adaptée pour une application particulière. Son comportement a été testé sur une large base et présente des résultats intéressants.

Introduction

L'interprétation d'images concerne de nombreuses applications, notamment la détection de cibles et leur reconnaissance, l'imagerie médicale ou la vidéo surveillance. Quelle que soit l'application concernée, la qualité de l'extraction de l'information conditionne les performances de l'algorithme. Pour chaque objet d'intérêt, la qualité de la localisation et la reconnaissance est très importante. De nombreux algorithmes ont été proposés dans la littérature (Cucchiara et al. (2003); Dalal et Triggs (2005); Jurie et Schmid (2004); Csurka et al. (2004)), mais il est encore difficile de comparer leurs performances.

Afin d'évaluer les algorithmes de détection et de reconnaissance d'objets, plusieurs compétitions ont vu le jour tels que le Pascal VOC Challenge (Everingham et al. (2008)) ou le projet Robin (D'Angelo et al. (2006)). Pour une vérité terrain donnée, ces compétitions utilisent des métriques afin d'évaluer et de comparer les résultats obtenus par différents algorithmes d'interprétation d'images. Si ces métriques font *a priori* appel à des caractérisations relevant du bon sens, aucune ne met en avant les mêmes caractéristiques. L'objectif de ces compétitions étant de comparer les algorithmes d'interprétation d'images en étudiant leur comportement global face à différents scénarios, il est primordial de disposer d'une métrique d'évaluation fiable.

De nombreuses métriques initialement proposées dans différents domaines peuvent se révéler utiles pour l'évaluation de résultats d'interprétation d'images. L'objectif de nos travaux est de mettre en concurrence ces métriques issues de contextes différents, afin de définir une métrique fiable d'évaluation de résultats d'interprétation prenant en compte la qualité de la détection, de la localisation et de la reconnaissance des objets. Dans l'exemple donné figure 1, nous aimerions ainsi pouvoir distinguer automatiquement le meilleur résultat d'interprétation.

Les métriques de la littérature trouvent leurs limitations dans leurs incapacités à évaluer la détection, la localisation et la reconnaissance en même temps. De plus, l'évaluation est souvent faite en moyenne sur une base importante. Il manque donc une métrique capable d'évaluer un résultat d'interprétation seul, en termes de localisation et de reconnaissance des objets d'intérêt. L'objectif de ce papier est de proposer une métrique permettant de combler ce manque.

Cet article est organisé de la façon suivante : la première section présente les méthodes de la littérature, puis la seconde section présente la métrique développée. La troisième section est consacrée aux résultats que nous avons obtenus avec cette métrique. Enfin, nous concluons et présentons quelques perspectives.

1 État de l'art

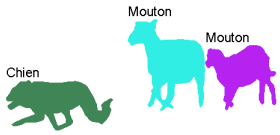
Nous présentons dans cette section un bref état de l'art des méthodes existantes d'évaluation supervisée, c'est-à-dire lorsque nous disposons de la vérité terrain, de la qualité de la localisation et de la reconnaissance d'objets dans une image.

1.1 Évaluation de la localisation

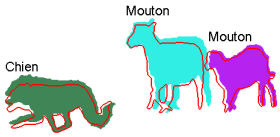
L'évaluation supervisée d'un résultat de localisation consiste à comparer deux images : la vérité terrain et le résultat de localisation. Plusieurs métriques ont été proposées pour cela (Basseville (1989); Wilson et al. (1997); Martin et al. (2001); Hafiane et al. (2007)), avec des objectifs initiaux comme l'évaluation de la segmentation, et se révèlent utilisables dans le cadre de l'évaluation de résultats de localisation. Cependant, l'existence d'un grand nombre de métriques indique clairement le manque d'une métrique connue faisant le consensus.

Par ailleurs, trois modes de représentation d'un résultat de localisation sont utilisés. Le plus simple est l'utilisation de boîtes englobantes, c'est-à-dire de rectangles présents dans l'image dont l'intérieur comprend l'objet recherché. Elles sont alors représentées par un ensemble de coordonnées dans l'image. La seconde méthode consiste à utiliser des images présentant les pixels frontières, ou contours, des objets à localiser. Les pixels de ces images prennent leurs valeurs dans $\{0, 1\}$. Les 0 représentent l'objet et le fond de l'image qui sont délimités par les

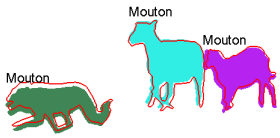
(a) Image originale



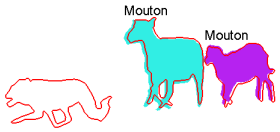
(b) Vérité terrain



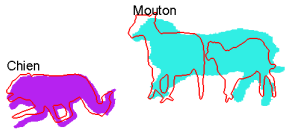
(c) Résultat d'interprétation 1



(d) Résultat d'interprétation 2



(e) Résultat d'interprétation 3



(f) Résultat d'interprétation 4

FIG. 1 – *Exemples de résultats d'interprétation d'images (image originale et vérité terrain issues d'Everingham et al. (2008)) : les lignes rouges correspondent à la vérité terrain*

pixels frontières de valeur 1. Enfin, il est possible d'utiliser des images présentant une région d'intérêt, ou un masque. Les pixels de ces images prennent également leurs valeurs dans $\{0, 1\}$, les pixels 0 représentent le fond tandis que les pixels 1 représentent l'objet. Les trois types de représentation d'un résultat de localisation sont présentés à la figure 2. Quelques exemples de métriques d'évaluation de la localisation sont donnés ci-dessous, pour chaque type de représentation d'un résultat de localisation.

Le projet Robin (D'Angelo et al. (2006)) visait à évaluer des résultats d'interprétation. Cependant, la localisation et la reconnaissance sont évaluées séparément. La localisation est évaluée à partir d'une représentation par boîtes englobantes par les trois métriques suivantes :

$$ROB_{loc}(BB_l, BB_{vt}) = \frac{2}{\pi} arctan(max(\frac{|x_l - x_{vt}|}{w_{vt}}, \frac{|y_l - y_{vt}|}{h_{vt}})) \quad (1)$$

(a) Boîte englobante (en rouge) (b) Frontière (en rouge) (c) Région

FIG. 2 – *Différentes représentations d'un résultat de localisation*

$$ROB_{com}(BB_l, BB_{vt}) = \frac{|\mathcal{A}_l - \mathcal{A}_{vt}|}{max(\mathcal{A}_l, \mathcal{A}_{vt})} \tag{2}$$

$$ROB_{cor}(BB_l, BB_{vt}) = \frac{2}{\pi} arctan(|\frac{h_l}{w_l} - \frac{h_{vt}}{w_{vt}}|) \tag{3}$$

où BB_l est la boîte englobante du résultat de localisation, x_l, y_l les coordonnées du centre de la boîte englobante, w_l, h_l la largeur et la hauteur de la boîte englobante et $\mathcal{A}_l$ son aire. Les valeurs indicées par vt représentent leurs équivalents pour la vérité terrain. Ces trois métriques évaluent des caractéristiques différentes des boîtes englobantes : ROB_{loc} évalue la position du centre de la boîte englobante, ROB_{com} évalue la taille de la boîte englobante et ROB_{cor} s'intéresse au rapport hauteur/largeur de la boîte englobante.

Pratt et al. (1978) ont proposé une mesure empirique (Figure of Merit) entre un résultat de localisation I_l et une référence I_{vt}, en utilisant une représentation d'un résultat de localisation par des frontières.

$$FOM(I_{vt}, I_l) = \frac{1}{MP} \sum_{k \in I_l^{Fr}} \frac{1}{1 + \alpha * d(k, I_{vt}^{Fr})^2} \tag{4}$$

avec MP correspondant à $\max(|I_{vt}^{Fr}|, |I_l^{Fr}|)$, I_{vt}^{Fr} l'ensemble des pixels frontières de la vérité terrain, $|I_{vt}^{Fr}|$ le cardinal de l'ensemble I_{vt}^{Fr}, α une constante de normalisation et $d(x, I) = \min_{y \in I} d(x, y)$. Un inconvénient de la mesure de Pratt est qu'elle n'est pas sensible aux erreurs de sous-détection alors qu'elle est sensible aux erreurs de sur-détection et de localisation. De plus, elle n'est pas sensible à la forme des zones erronées ce qui pose problème pour l'évaluation de la localisation d'un objet.

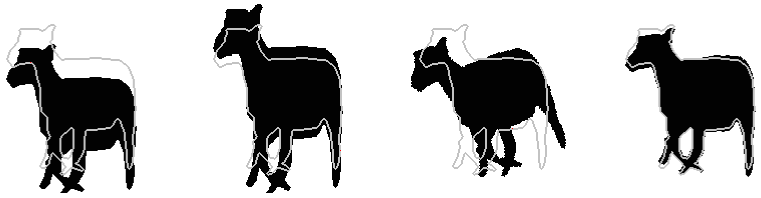
Le projet Pascal VOC Challenge (Everingham et al. (2005)) utilise une métrique simple, utilisant une représentation région, pour évaluer la localisation d'un seul objet :

$$PAS(I_{vt}, I_l) = \frac{|I_{vt}^{Re} \cap I_l^{Re}|}{|I_{vt}^{Re} \cup I_l^{Re}|} \tag{5}$$

où I_{vt}^{Re} correspond à l'ensemble des pixels appartenant à la région de l'objet d'intérêt dans l'image I_{vt}, $I_{vt}^{Re} \cap I_l^{Re}$ correspond à l'ensemble des pixels de l'objet correctement localisés et

$I_{vt}^{Re} \cup I_{l}^{Re}$ correspond à l'ensemble des pixels appartenant à l'objet ou reconnus comme tels. Cette métrique est comprise entre 0 et 1 et qualifie la qualité des pixels détectés de l'objet. Le résultat 1 correspond à un résultat optimal, c'est-à-dire lorsque $I_{vt}^{Re} \cap I_{l}^{Re} = I_{vt}^{Re} \cup I_{l}^{Re}$ soit $I_{vt}^{Re} = I_{l}^{Re}$.

Dans un article précédent (Hemery et al. (2010)), nous avons réalisé une étude comparative de 33 métriques d'évaluation de résultats de localisation, incluant les métriques présentées ci-dessus. Pour cela, nous avons créé une base de données synthétiques contenant un total de *118 080* résultats de localisation. Les données ont été générées en prenant en compte quatre altérations : la translation, la mise à l'échelle, la rotation et la perspective. La figure 3 présente les quatre altérations appliquées au même objet original.



(a) Translation (b) Mise à l'échelle (c) Rotation (d) Perspective

FIG. 3 – *Quatre images avec la même puissance d'altération (les lignes grises représentent les contours de l'objet original, la région noire représente l'objet altéré)*

Nous avons étudié les résultats d'évaluation au regard de quatre propriétés : la symétrie (une métrique doit pénaliser identiquement une altération dans une direction et dans celle opposée), la stricte monotonie (plus on altère un résultat de localisation par rapport à la vérité terrain, plus la métrique d'évaluation doit pénaliser cette altération), la continuité uniforme (le fait qu'il n'y ait pas de saut brusque dans l'évolution de la métrique d'évaluation) et la dépendance topologique (la métrique doit prendre en compte la forme et la taille des objets localisés). Notre conclusion était alors que les métriques exploitant une représentation basée région de l'objet permettent d'obtenir de meilleurs résultats. De plus, les métriques de Pascal (Everingham et al. (2008)) et de Martin et al. (2001) ont été jugées comme les plus pertinentes d'après notre étude.

1.2 Évaluation de la reconnaissance

Nous avons vu qu'il était possible de quantifier la qualité d'un résultat de localisation étant donnée une vérité terrain. Nous souhaitons faire de même avec un résultat de reconnaissance. Cependant, il n'est pas possible de calculer directement une distance sur les labels renvoyés par les algorithmes, c'est-à-dire sur les identifiants numériques ou les chaînes de caractères renvoyés. En effet, ces identifiants sont des variables qualitatives non ordonnées et calculer une distance entre ces variables n'aurait pas de sens. Naïvement, la seule méthode permettant de calculer une distance sur ces variables est de regarder si celles-ci sont égales, et donc de dire que la distance est de 0 si les identifiants sont identiques et 1 sinon. Cette quantification

reste peu précise et ne permet pas de pondérer une erreur entre classes.

Il est intéressant de remarquer qu'il est possible de calculer *a priori*, c'est à dire avant même d'avoir des résultats de reconnaissance à évaluer, l'ensemble des distances entre toutes les classes d'objets présentes dans la base de données utilisée pour l'application visée. En effet, le nombre de classes est limité et connu d'avance pour une application particulière. A titre d'exemple, nous souhaiterions que la distance entre les classes « chat » et « chien » soit plus petite que la distance entre les classes « voiture » et « chien ». L'ensemble de ces distances peut alors être stocké dans une matrice de distance MD ou matrice de similarité MS. Le problème est donc de calculer cette matrice.

Une façon de palier ce problème consiste à calculer la distance entre les descriptions des objets. La distance est alors dépendante de la méthode de représentation de l'objet. Par exemple, si l'objet est représenté par un histogramme, comme c'est le cas avec l'utilisation des méthodes se basant sur des sacs de mots, l'évaluation de la reconnaissance se fera en comparant les histogrammes des deux classes renvoyés par l'algorithme. Dans le cas d'une représentation des objets par un graphe, la distance d'édition peut être utilisée. Celle-ci peut être vue comme le coût d'édition du premier graphe G_1 pour le transformer en un graphe isomorphe au graphe G_2. Cette édition se fait en une succession d'éditions élémentaires, chacune ayant un coût individuel. La distance d'édition est alors la somme des coûts nécessaires à cette transformation. Cette distance se révèle compliquée à calculer, aussi les algorithmes de Munkres (1957) ou de Riesen et al. (2007) peuvent-ils être utilisés afin d'accéder à une approximation de cette distance.

Dans un article précédent (Hemery et al. (2009)), nous nous sommes intéressés à la représentation d'un objet par un nuage de points de descripteurs. Ce nuage de points est l'ensemble des points détectés automatiquement sur l'image. Chacun de ces points est ensuite caractérisé par un vecteur calculé sur son voisinage. Pour construire ce nuage de points, nous avons utilisé les descripteurs SIFT proposés par Lowe (2004), afin de détecter et caractériser les points d'intérêt de l'objet. Disposant de deux images représentant des objets, nous avions donc deux nuages de points, chacun associé à un objet. Nous avons ensuite apparié les points du premier objet avec ceux du second lorsque cela était possible (en considérant la valeur du descripteur), puis nous avons défini une mesure de similarité basée sur le nombre d'appariements réalisés. Il est alors possible d'obtenir une mesure de similarité entre plusieurs classes d'objets en prenant les scores de similarité moyens obtenus sur une grande base d'apprentissage.

Enfin, on pourra construire la matrice MD ou MS (si l'on utilise une mesure de similarité), soit en la remplissant manuellement de façon subjective, soit en utilisant une connaissance a priori comme une taxonomie ou une hiérarchie.

1.3 Évaluation de l'interprétation

Toutes les métriques d'évaluation proposées dans la littérature s'intéressent soit aux aspects de localisation soit aux aspects de reconnaissance des objets d'intérêt. Une métrique permettant d'évaluer un résultat d'interprétation, c'est-à-dire prenant en compte, en même temps, ces deux aspects n'existe pas à l'heure actuelle. Ceci est la principale contribution de ce papier.

2 Métrique développée

Comme nous pouvons le voir à la figure 4, la méthode que nous avons développée est composée de quatre étapes : (I) une mise en correspondance, (II) une évaluation locale pour chaque objet mis en correspondance, (III) une prise en compte de la sous et de la sur-détection pour chaque objet non mis en correspondance, et (IV) un calcul du score global. Ces étapes sont décrites dans les paragraphes suivants.

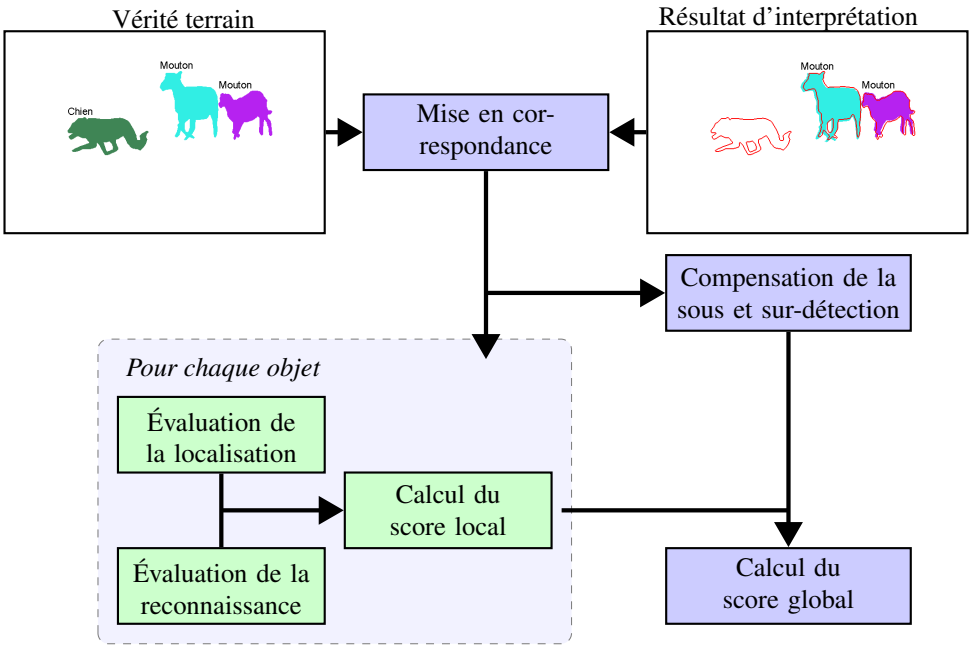


FIG. 4 – *Schéma de l'évaluation globale d'un résultat d'interprétation*

2.1 La mise en correspondance

La mise en correspondance des objets permet de déterminer quels objets de la vérité terrain correspondent aux objets détectés par l'algorithme d'interprétation. Cette étape est nécessaire pour les deux étapes suivantes : le calcul des scores locaux et la compensation de la sous et sur-détection. En effet, les objets sur-détectés correspondent à des objets présents dans le résultat d'interprétation qui ne correspondent à aucun objet de la vérité terrain. Les objets sous-détectés sont les objets présents dans la vérité terrain mais qui sont absents du résultat d'interprétation. Afin de faire cette mise en correspondance, nous calculons une matrice de recouvrement, de manière similaire aux travaux présentés par Phillips et Chhabra (1999). Chaque ligne de la matrice correspond à un objet présent dans la vérité terrain, tandis que les colonnes correspondent aux objets présents dans le résultat d'interprétation. Dans la cellule (u, v), nous indiquons le

recouvrement de l'objet u de la vérité terrain et de l'objet v du résultat d'interprétation. Le recouvrement est calculé avec la métrique PAS présentée dans la section précédente. La figure 5 présente l'ensemble $I_{vt}^{Re} \cap I_i^{Re}$ en vert tandis que l'ensemble $I_{vt}^{Re} \cup I_i^{Re}$ est la réunion des régions verte, bleue et rouge.

(a) Vérité terrain (b) Résultat de localisation (c) Superposition des deux régions

FIG. 5 – *La métrique PAS correspond à l'ensemble gris foncé de pixels en commun divisé par l'ensemble des pixels localisés*

A partir de cette matrice, nous avons choisi d'implémenter deux méthodes pour la mise en correspondance. La première méthode est en « un pour un » et consiste à assigner un seul objet de la vérité terrain à un seul objet du résultat d'interprétation. Cette méthode est également utilisée dans Everingham et al. (2008). Nous utilisons l'algorithme hongrois (Munkres (1957)) afin de calculer cette mise en correspondance. La deuxième méthode, dite « multiple », permet que chaque objet de la scène puisse être assigné à plusieurs objets dans la vérité terrain ou le résultat d'interprétation. Pour cela, nous faisons simplement un seuillage sur la matrice de recouvrement. Cela permet de prendre en compte, par exemple, le fait qu'un groupe d'objets soit reconnu comme un seul objet par l'algorithme. Cette méthode est utilisée dans Phillips et Chhabra (1999) ou Wolf et Jolion (2006). Après cette mise en correspondance, nous obtenons une matrice de correspondance, où un 1 dans une cellule (u, v) indique que l'objet u de la vérité terrain est mis en correspondance avec l'objet v du résultat d'interprétation. Par défaut, la méthode utilisée est la méthode « multiple » avec un seuil de $0,2$. Ces paramètres sont modifiables par l'utilisateur afin d'adapter la méthode d'évaluation en fonction de l'application visée.

2.2 Le calcul du score local

Pour chaque objet mis en correspondance, c'est-à-dire les cases (u, v) contenant un 1 dans la matrice de correspondance, nous calculons un score, dit « local », correspondant à l'évaluation d'un seul objet dans une scène en contenant potentiellement plusieurs. Il est calculé à partir de la qualité de la localisation et de la reconnaissance de l'objet. Le score local de localisation S_{loc} est une version simplifiée pour un objet des métriques de Martin et al. (2001) qui ont montré leurs bonnes performances dans nos travaux précédents :

$$S_{loc}(I_{vt}, I_i, u, v) = \min\left(\frac{|I_{vt\setminus i}^{Re(u)}|}{|I_{vt}^{Re(u)}|}, \frac{|I_{i\setminus vt}^{Re(v)}|}{|I_i^{Re(v)}|}\right) \quad (6)$$

avec $|I_{vt}^{Re(u)}|$ le nombre de pixels de l'objet u présents dans la vérité terrain et $|I_{vt\setminus i}^{Re(u)}|$ le nombre de pixels de l'objet u présents dans la vérité terrain, mais pas dans le résultat d'interprétation. Ce score est compris entre 0 et 1, 0 indiquant un recouvrement parfait des deux objets (parfaite localisation).

Nous calculons ensuite un score pour la qualité de la reconnaissance. Pour cela, nous utilisons une matrice de distance entre les classes d'objets présentes dans la base de données. L'utilisateur doit fournir cette matrice de distance qu'il aura précédemment calculée ou, éventuellement, construite manuellement, comme nous l'avons vu dans la section 1.2. Si l'utilisateur ne fournit pas de matrice de distance, une matrice carrée contenant autant de lignes qu'il y a de classes dans la base de données est construite par défaut. Elle ne contient que des 1 avec des 0 sur la diagonale, soit la matrice $(1 - \delta_{i,j})$, avec $\delta_{i,j}$ le symbole de Kronecker. De plus, l'algorithme d'interprétation d'images peut fournir un indice de confiance pour chaque objet détecté. Le score est alors le suivant :

$$S_{rec}(I_{vt}, I_i, u, v, \mu) = MD(cl(u), cl(v)) * ind(cl(u), cl(v), \mu) \tag{7}$$

avec

$$ind(cl(u), cl(v), \mu) = \begin{cases} \frac{1-\mu}{2} & \text{si } cl(u) = cl(v) \\ \frac{1+\mu}{2} & \text{sinon} \end{cases} \tag{8}$$

et μ l'indice de confiance accordé au résultat de reconnaissance, MD la matrice des distances entre les classes et $cl(u)$ la classe de l'objet u.

Suite à cela, nous calculons un score d'interprétation, qui est la combinaison de scores de localisation et de reconnaissance :

$$S(u, v, \mu) = \alpha * S_{loc}(I_{vt}, I_i, u, v) + (1 - \alpha) * S_{rec}(I_{vt}, I_i, u, v, \mu) \tag{9}$$

Par défaut, la valeur du paramètre α est de $0,8$, mais l'utilisateur peut choisir de le modifier afin de donner plus de poids à la reconnaissance ou à la localisation. Nous avons choisi la valeur de $0,8$ afin de prendre davantage en considération la localisation que la reconnaissance étant donnée que la pénalisation par défaut de la reconnaissance est plus importante que celle de la localisation.

2.3 La compensation

Après avoir calculé un score pour chaque objet mis en correspondance, nous regardons les objets sous et sur-détectés. La sous-détection correspond à des lignes de la matrice de correspondance qui n'ont pas été associées à un objet du résultat d'interprétation, donc n'ayant pas de 1. De même, les objets sur-détectés vont correspondre à des colonnes de la matrice de correspondance qui n'ont été associées à aucun objet de la vérité terrain. Nous commençons par prendre en compte la sous-détection. Pour cela, on recherche la première ligne u sans association, puis pour cette ligne, on recherche la première colonne v sans association. On associe alors l'objet u à l'objet v dans la matrice de correspondance. Dans la matrice des scores locaux, le score de cette association est mis à 1 ce qui correspond à un mauvais score

d'interprétation. On recommence ensuite jusqu'à ce que toutes les lignes soient associées. Pour la sur-détection, on fait le même travail en échangeant les lignes et les colonnes.

2.4 Le calcul de score global

Le score global est calculé à partir de la matrice des scores locaux, le score global étant la moyenne des scores locaux.

2.5 Illustration

Nous avons appliqué notre méthode d'évaluation, avec les paramètres par défaut, à un résultat d'interprétation de l'image présentée à la figure 6. À la figure 7, nous pouvons voir

FIG. 6 – *Une image originale de la base Pascal VOC challenge 2007*

chaque étape de l'évaluation. La vérité terrain présente sept objets : les quatre premiers sont de la classe « personne », suivis d'un objet de la classe « bus », puis « avion » et enfin un objet de la classe « voiture » qui est peu visible. Le résultat d'interprétation présente quatre objets : le premier de la classe « camion », suivi d'un objet de la classe « avion » puis de deux objets de la classe « personne ».

La matrice intitulé « Recouvrement » présente quatre colonnes correspondant aux quatre objets présents dans le résultat d'interprétation, les sept lignes correspondant quant à elles aux objets présents dans la vérité terrain. Nous pouvons constater que l'avion et le bus ont été bien localisés, mais que le bus est reconnu comme un camion. Les quatre personnes ont bien été reconnues mais mal localisées. En effet, un seul objet a été détecté à la place de trois. Enfin, la voiture n'a pas été détectée du tout.

La première étape consiste à mettre en correspondance les objets de la vérité terrain et ceux du résultat d'interprétation. La matrice de recouvrement présente le résultat de la métrique PAS pour chaque couple (u, v) d'objets. Ainsi, pour le bus, qui est l'objet 5 de la vérité terrain et l'objet 1 du résultat d'interprétation, le recouvrement des masques est important. Cela amène un score de recouvrement de 0,941. Ce score étant supérieur au seuil, ces deux masques sont mis en correspondance comme on peut le voir dans la matrice de correspondance. Il en va

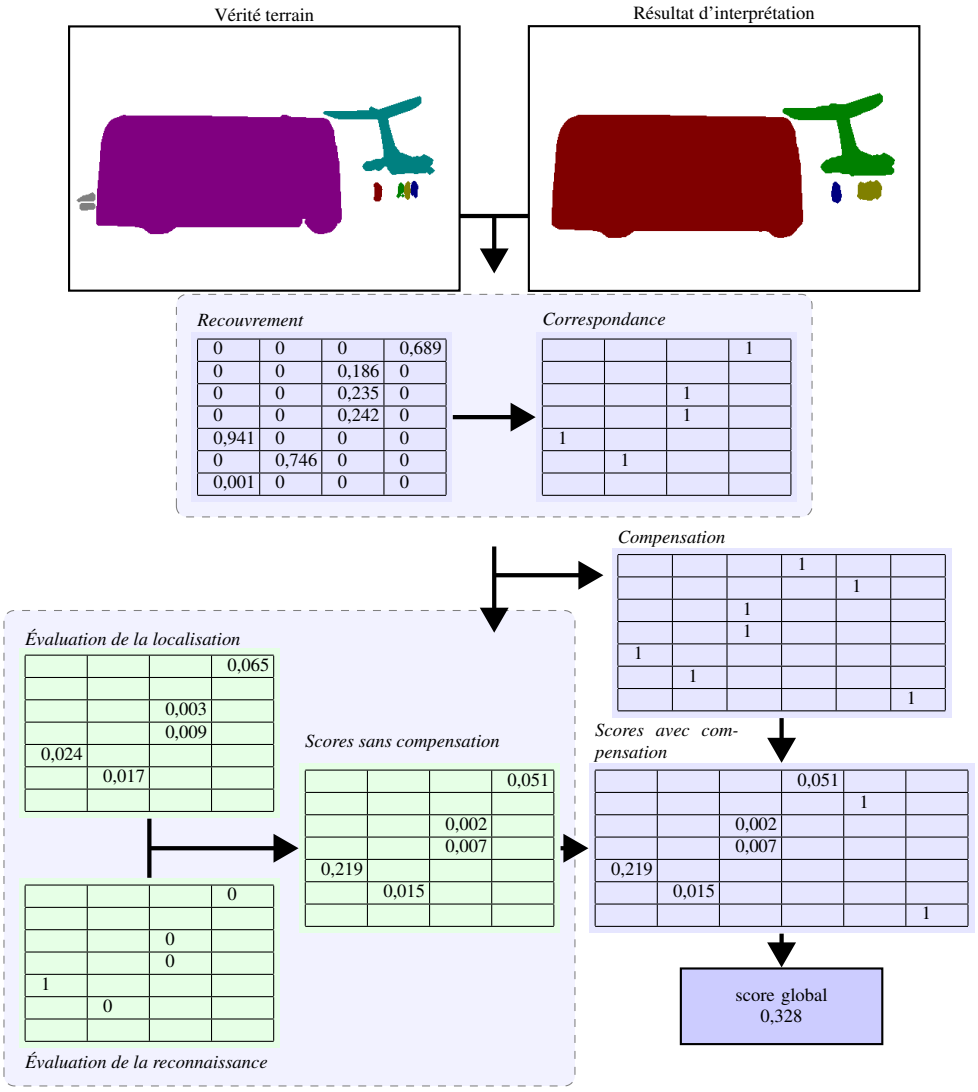


FIG. 7 – *Exemple d'évaluation globale d'un résultat d'interprétation*

de même pour l'avion ainsi qu'une personne qui sont très clairement mis en correspondance. Le groupe de trois personnes correspond aux objets 2, 3 et 4 de la vérité terrain et à l'objet 3 du résultat d'interprétation. Le score de recouvrement est donc moins important avec 0,235 et 0,242 pour 2 objets et 0,186 pour le dernier. Le seuil étant de 0,2, seulement 2 objets de la vérité terrain sont mis en correspondance avec l'objet du résultat d'interprétation.

La seconde étape consiste à calculer les scores locaux obtenus pour chaque objet mis en correspondance. Pour cela, une matrice des scores de localisation est calculée, ainsi qu'une ma-

trice des scores de reconnaissance. Pour la localisation, on peut voir que le groupe d'individus est bien localisé avec des scores de localisation inférieurs à 0,01 tandis que l'autre individu est moins bien localisé avec un score de 0,065, ce qui reste un bon score de localisation. L'avion et le bus sont bien localisés avec des scores de localisation de 0,017 et 0,024. Pour la reconnaissance, nous pouvons voir qu'à l'exception du bus, les objets sont bien reconnus ce qui explique qu'ils aient un score de 0. Le bus étant reconnu comme un camion, son score est de 1. L'utilisation d'une matrice de similarité à définir en fonction du contexte permettrait de réduire ce score étant donné que ces deux objets sont assez similaires. Une matrice de scores locaux est ensuite calculée comme la combinaison des deux matrices précédentes. Nous pouvons voir que le score du bus est fortement impacté par la mauvaise reconnaissance.

La troisième étape est la compensation. On part pour cela de la matrice de correspondance afin d'identifier les lignes et/ou les colonnes n'ayant pas été affectées. Nous pouvons voir que toutes les colonnes contiennent au moins un 1, ce qui signifie que tous les objets du résultat d'interprétation ont été affectés à au moins un objet de la vérité terrain, et qu'il n'y a pas de sur-détection. Cependant, les lignes 2 et 7 sont vides, les objets correspondant dans la vérité terrain n'ont pas été convenablement détectés. Cette sous-détection est alors compensée en ajoutant des scores 1 dans les lignes correspondantes. Des colonnes sont ajoutées afin de ne pas affecter ces objets à des objets du résultat d'interprétation déjà correctement détectés.

La dernière étape consiste à calculer la matrice des scores locaux en prenant en compte la compensation. Cette matrice présente les scores locaux calculés précédemment en ajoutant les scores provenant de la compensation. A partir de cette matrice, on calcule le score moyen ce qui nous donne le score global. Dans notre cas, on calcule la moyenne des 7 scores obtenus : les 5 provenant de la mise en correspondance et les 2 provenant de la compensation. Le score final obtenu dans le cas étudié est de 0,328 (le score optimal correspond à la valeur nulle). Ce score est assez élevé car l'absence de détection de deux objets est très pénalisante : les objets manquants contribuent à hauteur de 0,285 et les objets présents à hauteur de 0,043.

3 Validation de la méthode

3.1 Protocole expérimental

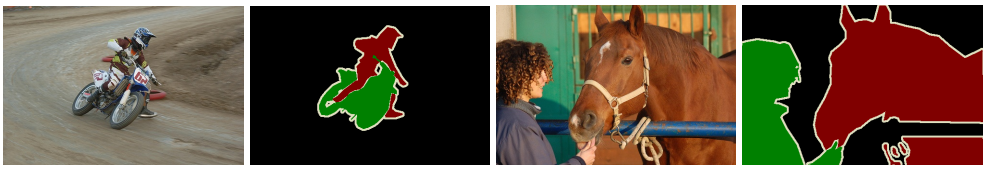
Nous avons testé notre méthode d'évaluation globale sur une grande base de données dont les vérités terrain contiennent à la fois la localisation par des masques ainsi qu'une classe associée à chaque objet. Nous avons ensuite appliqué sur les objets issus de ces vérités terrain des altérations, puis nous avons étudié le comportement de notre méthode en fonction des différentes altérations. La suite de cette section présente la base de données ainsi que les altérations que nous avons appliquées aux vérités terrain.

3.1.1 Base de données

Nous avons utilisé la base de données du Pascal VOC challenge 2008 [1] (Everingham et al. (2008)). Parmi cette base de données, plusieurs ensembles sont disponibles, chacun correspondant à un type d'algorithme. Nous avons utilisé l'ensemble « Segmentation Taster Set ». Cet

1. http://www.pascal-network.org/challenges/VOC/databases.html

ensemble d'images présente des vérités terrain dont la localisation est un masque, ce qui correspond à notre algorithme d'évaluation. Nous pouvons voir quelques exemples d'images issues de cet ensemble à la figure 8. Dans la table 1, nous référençons le nombre d'images contenant



(a) Image (b) Vérité terrain (c) Image (d) Vérité terrain

FIG. 8 – *Images provenant de l'ensemble « Segmentation Taster Set »*

un nombre N d'objets. Nous pouvons voir qu'une majorité des images contient seulement un ou deux objets. Sur les 1022 images disponibles dans la base, 1002 ont entre 1 et 8 objets. Les 20 images restantes ont entre 9 et 21 objets. Comme il n'y avait pas au minimum 10 images contenant N (N>8) objets, ces images ont été rejetées. La base de données finale représente donc un total de *2 134* objets que nous avons altérés.

Nombre d'objets	1	2	3	4	5	6	7	8
Nombre d'images	498	237	106	61	40	32	16	12

TAB. 1 – *Nombres d'images utilisées en fonction du nombre d'objets contenus*

Cette base contient en tout 20 classes différentes : « avion », « bicyclette », « oiseau », « bateau », « bouteille », « bus », « voiture », « chat », « chaise », « vache », « table à manger », « chien », « cheval », « moto », « personne », « plante en pot », « mouton », « sofa », « train » et « télévision ». Nous avons ordonné ces classes selon les catégories de la base Caltech256 (Griffin et al. (2007)), présentée à la figure 9, afin de créer la matrice de distance.

3.1.2 Altérations

Nous avons tout d'abord considéré les mêmes altérations que lors de l'étude comparative réalisée sur la localisation (Hemery et al. (2010)) : la translation, la mise à l'échelle, la rotation et enfin la perspective. Pour chacune de ces altérations, nous avons utilisé un paramètre d'altération allant de 1 à 20 dans deux directions différentes : horizontale et verticale pour les altérations de translation, mise à l'échelle et perspective, sens horaire et antihoraire pour la rotation. Cela a amené à créer 160 altérations pour chaque objet, soit un total de *341 440* altérations considérées. Nous avons ensuite simulé des erreurs de reconnaissance en altérant la classe des objets. Pour cela, nous avons remplacé la classe de 1 à tous les objets présents dans l'image par la classe « $Autre$ ». De même, nous avons regardé l'effet de la sous et sur-détection d'objets sur le comportement de notre méthode d'évaluation. Pour cela, nous avons ajouté ou supprimé de 1 à 8 objets, de sorte qu'ils ne soient en correspondance avec aucun autre objet présent dans l'image.

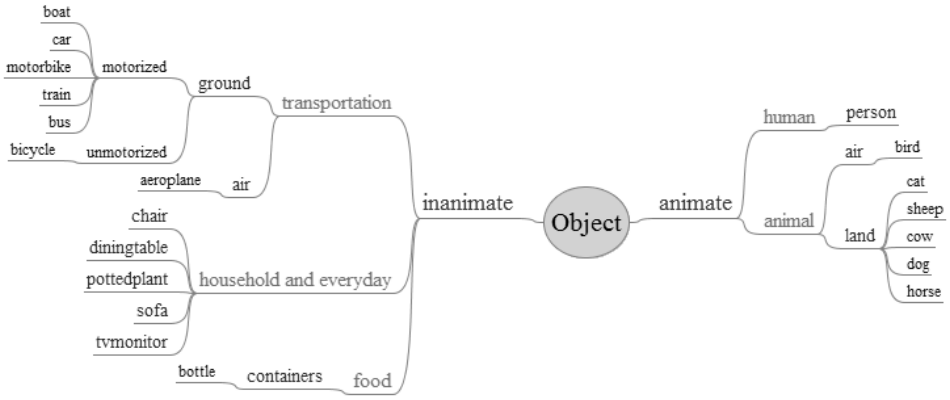


FIG. 9 – *Classes de la base Pascal 2008 représentées en utilisant la taxonomie de la base Caltech256*

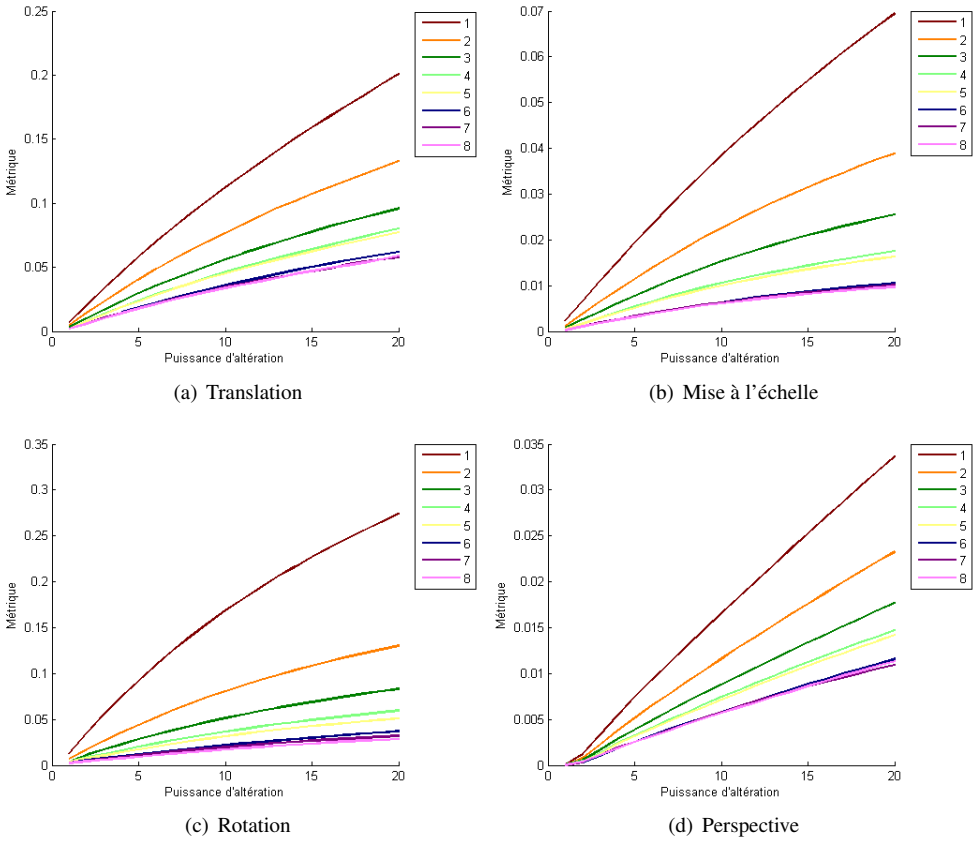
3.1.3 Paramétrage

Nous nous sommes également intéressés à l'évolution de la métrique en fonction des différents paramètres de celle-ci. Nous avons alors regardé l'effet du choix de mise en correspondance ainsi que l'effet du seuil sur la méthode de mise en correspondance « multiple ». Nous avons donc comparé la mise en correspondance « un pour un » et « multiple », avec un seuil valant 0,2, 0,3, 0,4 et 0,5. Nous avons également regardé les effets de l'utilisation d'un indice de confiance pour la reconnaissance, ainsi que l'utilisation d'une matrice de similarité entre les différentes classes d'objets.

3.2 Résultats expérimentaux

3.2.1 Localisation

La figure 10 présente l'évolution moyenne de la métrique globale en fonction de différentes altérations de la localisation et selon le nombre d'objets présents dans la vérité terrain. Chaque courbe présente l'évolution de la métrique en fonction de la puissance d'altération de la localisation. La métrique est ici présentée avec les valeurs par défaut. Nous remarquons tout d'abord que plus le nombre d'objets dans la vérité terrain augmente et moins le critère est pénalisant. Ceci est normal puisque le score global est la moyenne des scores locaux. Ce résultat est donc correct et en adéquation avec la manière dont nous avons développé la métrique. Nous pouvons voir également que les propriétés définies dans la section 1.1 sont respectées. Quelle que soit l'altération considérée ou le nombre d'objets, les courbes sont uniformément régulières et strictement monotones. Nous pouvons voir que la métrique a bien la propriété de séparabilité également. De plus, bien que cela ne soit pas visible sur les courbes, la métrique est symétrique. Nous pouvons également remarquer que la translation et la rotation sont les deux altérations les plus pénalisées (à puissance d'altération égale), suivies par le changement d'échelle puis le changement de perspective. Ceci semble correct au regard de la figure 3, où toutes les images



(a) Translation (b) Mise à l'échelle

(c) Rotation (d) Perspective

FIG. 10 – *Résultats concernant la localisation*

ont été altérées avec le même paramètre de 20. La métrique évolue donc correctement en ce qui concerne la localisation.

3.2.2 Reconnaissance

Concernant la reconnaissance, nous avons étudié l'évolution de la métrique en fonction du nombre d'objets dont nous avons altéré la classe. Étant donné que nous travaillons avec les paramètres par défaut, la classe affectée à l'objet altéré n'a pas d'importance sur le résultat. Nous pouvons voir à la figure 11 l'évolution de la métrique globale en fonction du nombre d'objets altérés, les différentes courbes représentant le nombre d'objets dans la vérité terrain. Nous pouvons remarquer que la métrique évolue correctement puisque la pénalisation maximale, d'une valeur de 0,2 ($1-\alpha$, le paramètre α étant fixé par défaut à 0,8), est atteinte lorsque tous les objets de la vérité terrain ont été altérés.

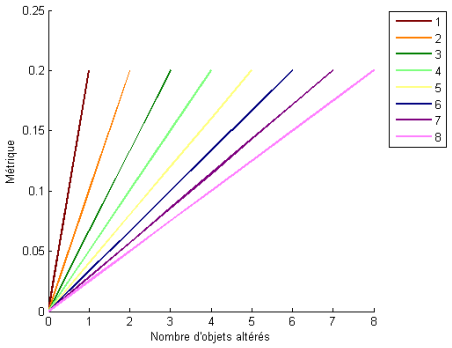


FIG. 11 – *Résultats concernant la reconnaissance*

3.2.3 Sous et sur-détection

Nous avons ensuite étudié l'effet de la sous et sur-détection sur l'évolution de la métrique globale. Nous avons étudié l'évolution de celle-ci, avec les paramètres par défaut, en fonction du nombre d'objets supprimés de la vérité terrain ou bien ajoutés. La figure 12 présente l'évolution de la métrique en fonction du nombre d'objets altérés (les nombres négatifs indiquent les objets supprimés et les positifs les objets ajoutés) où chaque courbe correspond à un nombre d'objets dans la vérité terrain. Nous pouvons voir que ces situations sont correctement gérées et que la métrique est toujours plus pénalisante lorsque le nombre d'objets dans la vérité terrain est faible. Nous pouvons noter que la sous-détection est légèrement plus pénalisée que la sur-détection. Tout cela montre que la métrique, lorsqu'elle est utilisée avec le paramétrage par défaut, donne de bons résultats. Cependant, il peut être intéressant de la paramétrer selon une application visée. La suite de cette section présente l'influence du paramétrage sur le comportement de cette métrique.

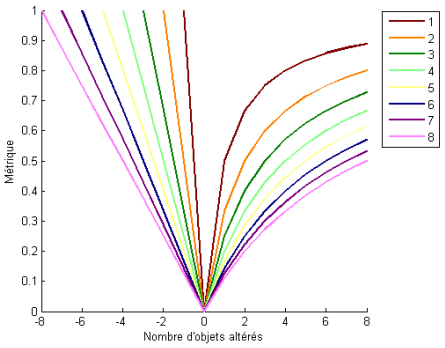
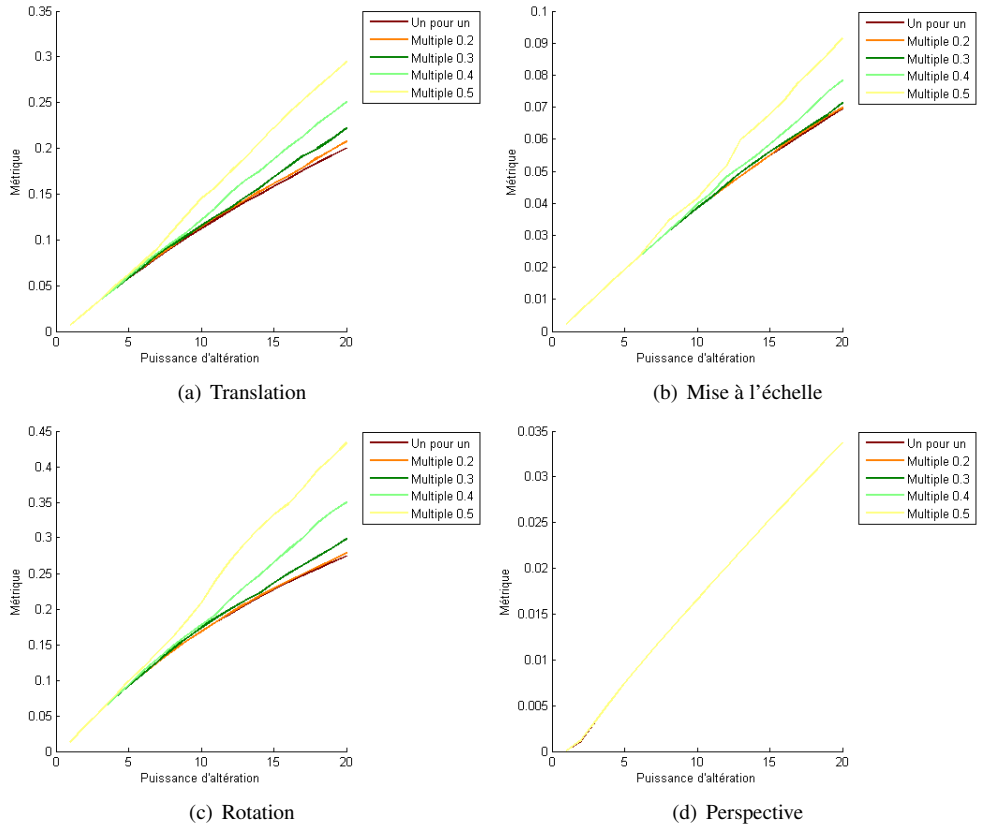


FIG. 12 – *Résultats concernant la sous et sur-détection*

3.2.4 Paramétrage

Mode de mise en correspondance Le premier paramètre que nous avons considéré est le mode de mise en correspondance, ainsi que le seuil de mise en correspondance. Pour cela, nous nous sommes intéressés à l'évolution de notre métrique d'évaluation en fonction des quatre altérations de la localisation. Nous nous sommes limités à l'étude des vérités terrain contenant un seul objet. Chaque courbe de la figure 13 représente un paramétrage différent de la métrique. Nous pouvons remarquer sur les résultats obtenus que le paramétrage « multiple » est plus pénalisant que le paramétrage « un pour un ». De plus, plus le seuil est élevé, et plus les altérations sont pénalisées. Cela s'explique par le fait qu'il n'y ait qu'un seul objet par vérité terrain. Ainsi, la méthode « un pour un » associe toujours l'objet altéré à l'objet de la vérité terrain, tandis que la méthode « multiple » ne l'associera plus à partir d'un certain degré d'altération, cela se produisant d'autant plus rapidement que le seuil est élevé. Notons également que cela se produit plus rapidement avec les altérations de rotation et de translation qu'avec l'altération de mise à l'échelle. Enfin, l'altération de perspective n'est pas assez importante pour que cela se produise.



(a) Translation (b) Mise à l'échelle

(c) Rotation (d) Perspective

FIG. 13 – *Résultats concernant l'effet du paramétrage sur l'évaluation de la localisation*

Indice de confiance Nous avons ensuite analysé l'effet de l'indice de confiance sur le résultat de reconnaissance. Pour cela, la figure 14 présente la valeur de la variable ind en fonction de l'indice de confiance μ, la courbe verte dans le cas où la classe est bien reconnue, la rouge dans le cas d'une erreur. Nous voyons clairement que plus la confiance accordée est importante, plus le score de reconnaissance est impacté car la différence entre la courbe rouge et la courbe verte augmente. Dans le cas d'une bonne reconnaissance, une augmentation de la confiance permet de diminuer la valeur de l'indice multiplicateur et donc du score de reconnaissance. Le score de reconnaissance étant plus bas, le score global le devient également, signe d'une meilleure interprétation. Au contraire, dans le cas d'une mauvaise reconnaissance, l'indice multiplicateur augmente avec la valeur de l'indice de confiance. Ainsi, plus la confiance est élevée dans un mauvais résultat, plus le score augmente, signe d'une mauvaise interprétation. Il est important de remarquer que l'indice multiplicateur est toujours plus grand dans le cas d'une mauvaise reconnaissance que dans le cas d'une bonne. Ainsi, quelle que soit la confiance accordée à un résultat, une mauvaise reconnaissance sera toujours plus pénalisée qu'une bonne reconnaissance.

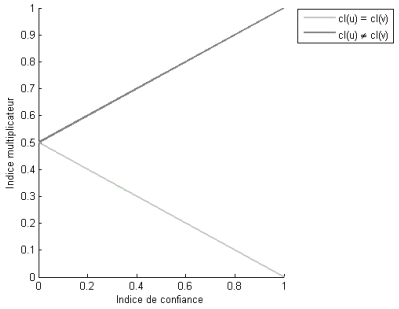
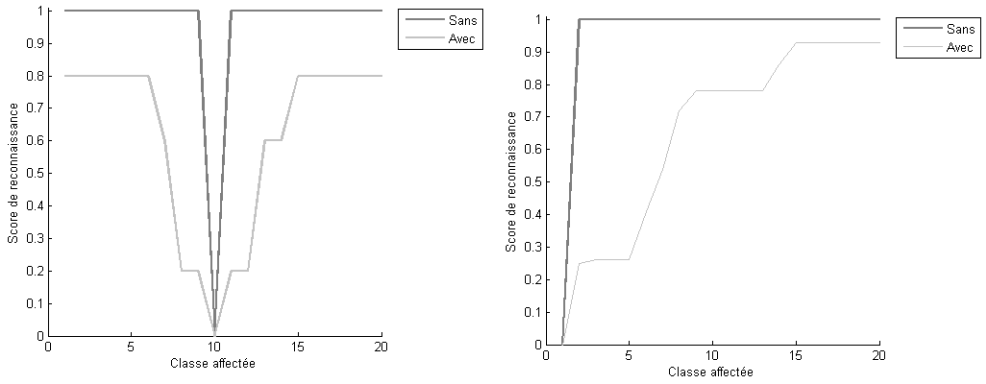


FIG. 14 – *Valeur de l'indice multiplicateur en fonction de l'indice de confiance*

Pondération de la reconnaissance Nous nous sommes enfin intéressés à l'effet de l'utilisation d'une matrice des distances entre classes sur les résultats de reconnaissance. Pour cela, nous avons calculé la matrice des distances à partir de la taxonomie créée à la figure 9. La distance entre deux classes étant égale à la distance entre leurs noeuds dans la taxonomie. La figure 15 nous montre l'évolution du score de reconnaissance en fonction de l'utilisation ou non d'une matrice des distances. La figure 15 (a) présente l'évolution du score de reconnaissance pour la classe 10 « plante en pot » en fonction de la classe assignée. On peut voir que l'utilisation d'une matrice de pondération de l'erreur de reconnaissance permet d'avoir un score évoluant plus finement. La figure 15 (b) présente l'évolution du score de reconnaissance en moyenne, pour les 20 classes, en fonction de la classe assignée. Nous avons trié les résultats afin de présenter l'évolution du score en fonction de la classe affectée, de la plus proche à la plus lointaine. Le score est mieux évalué avec une matrice de distance que sans.



(a) Résultat pour la classe 10 « plante en pot »

(b) Résultat moyen obtenu pour les 20 classes

FIG. 15 – *Valeur du score de reconnaissance avec et sans matrice de distance*

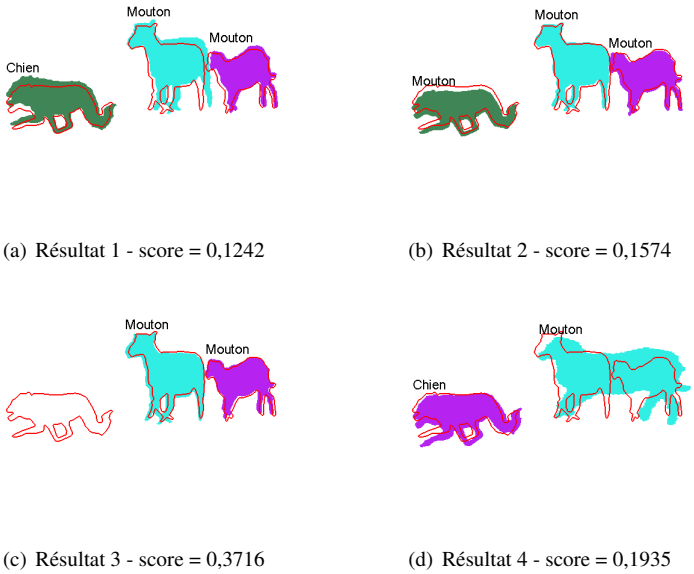
3.3 Discussion

Les résultats que nous avons obtenus avec la méthode proposée sont satisfaisants. Nous avons vu que la métrique permet de bien prendre en compte : (i) une mauvaise localisation, (ii) une mauvaise reconnaissance et (iii) une mauvaise détection. Il est à noter que les altérations sont pénalisées dans l'ordre d'importance suivant : d'abord la mauvaise détection, ensuite la reconnaissance puis la localisation. De même, parmi les altérations possibles de localisation, les problèmes de rotation et de translation sont pénalisés en priorité, suivis des problèmes de mise à l'échelle et enfin des problèmes de perspective.

Il est intéressant de remarquer que les cas des sur- et sous-détection sont traités à la fois par la mise en correspondance et la compensation. La mise en correspondance permet de gérer les cas où plusieurs objets sont détectés là où la vérité terrain n'en présente qu'un (et réciproquement), ce qui a peu d'impact sur la note finale. La compensation permet quant à elle de gérer les cas où des objets sont détectés là où aucun objet n'est présent dans la vérité terrain, ce qui a un fort impact sur la note finale. Nous avons également vu que la méthode est paramétrable et que cela influe sur les résultats d'évaluation. Le mode de mise en correspondance « multiple » permet notamment de rendre la méthode d'évaluation plus sévère en augmentant le seuil. L'ajout de l'indice de confiance renvoyé par l'algorithme ainsi que l'utilisation de matrice de similarité permettent d'avoir un résultat d'évaluation plus fin.

Si nous reprenons les résultats d'interprétation de la figure 1 et que nous les évaluons, nous obtenons les résultats présentés à la figure 16. Nous pouvons voir que le résultat (c) est le moins bon puisqu'il manque un objet. Suit le résultat (d), pour lequel il manque également un objet. Cependant, un objet du résultat d'interprétation chevauche deux objets de la vérité terrain, ce qui permet d'avoir une évaluation moins pénalisante que pour le résultat (c). Les résultats (a) et (b) obtiennent de meilleurs scores d'évaluation. Le premier ne contenant que des erreurs de localisation et aucune erreur de reconnaissance, il est mieux noté que le second. Des paramètres différents de ceux proposés par défaut permettraient d'avoir des résultats différents. Ainsi, avec

une valeur différente du paramètre α, il serait possible de mettre en avant le résultat (d) qui ne présente aucune erreur de reconnaissance, peu d'erreur de localisation mais simplement un problème de détection. La métrique peut donc être adaptée pour des applications présentant des enjeux spécifiques.



(a) Résultat 1 - score = 0,1242

(b) Résultat 2 - score = 0,1574

(c) Résultat 3 - score = 0,3716

(d) Résultat 4 - score = 0,1935

FIG. 16 – *Exemples d'évaluation de résultats d'interprétation sur une scène (image originale tirée de Everingham et al. (2008)). Le résultat 1 est jugé comme le meilleur par la métrique proposée si l'on utilise le paramétrage par défaut*

Enfin, il convient de remarquer que la métrique compare et donne une distance entre deux résultats d'interprétation. Pour effectuer une bonne évaluation d'un algorithme d'interprétation, l'un des deux résultats est en réalité une vérité terrain, c'est-à-dire un résultat d'interprétation idéale, fourni par des experts. Il est donc primordial d'avoir une bonne vérité terrain pour effectuer une bonne évaluation d'un algorithme. Cependant, la métrique présentée dans cet article peut également être utilisée pour comparer deux résultats d'interprétation provenant de deux algorithmes différents. Le résultat permettra alors de conclure si les deux algorithmes fournissent des résultats similaires, indépendamment de leur qualité.

4 Conclusions et perspectives

Nous avons travaillé sur la création d'une métrique permettant l'évaluation d'un résultat d'interprétation. Cette métrique permet de prendre en compte à la fois les informations concernant la localisation et la reconnaissance des objets dans la scène. Cette métrique se base sur une mise en correspondance, le calcul d'un score local pour chaque objet mis en correspondance, puis le calcul d'un score global prenant en compte la sous et sur-détection. Nous avons créé cette métrique afin qu'elle soit paramétrable pour s'adapter à une application visée. Nous

avons vu que la méthode de mise en correspondance peut être modifiée. Il en va de même de l'importance de la reconnaissance par rapport à la localisation. Enfin, l'utilisation d'une matrice de distance, créée automatiquement ou manuellement, permet de grandement améliorer les performances de la méthode d'évaluation. Les résultats obtenus par notre méthode d'évaluation correspondent aux objectifs que nous nous étions fixés. Nous avons vu qu'elle pénalise correctement les différentes altérations possibles et permet donc de comparer plusieurs résultats d'interprétation de façon à faire ressortir le meilleur.

Les perspectives concernent en premier lieu la création automatique de matrice de distance depuis une base de données. Une autre perspective est de faire une étude subjective permettant de comparer les résultats d'évaluation de la métrique proposée et les résultats d'évaluations réalisées par des humains. Ainsi, nous serons en mesure de paramétrer la métrique afin qu'elle reproduise au mieux un comportement humain. Enfin, la dernière perspective est d'utiliser la métrique sur diverses applications pratiques d'interprétation d'images.

Références

Basseville, M. (1989). Distance measures for signal processing and pattern recognition. *Signal Processing 18*(4), 349–369.

Csurka, G., C. R. Dance, L. Fan, J. Willamowski, et C. Bray (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision (ECCV)*, pp. 1–22.

Cucchiara, R., C. Grana, M. Piccardi, et A. Prati (2003). Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 25*(10), 1337–1342.

Dalal, N. et B. Triggs (2005). Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) 1*, 886–893.

D'Angelo, E., S. Herbin, et M. Ratiéville (2006). Robin challenge evaluation principles and metrics. http ://robin.inrialpes.fr.

Everingham, M., L. Van Gool, C. K. I. Williams, J. Winn, et A. Zisserman (2008). The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http ://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

Everingham, M., A. Zisserman, C. Williams, L. Van Gool, M.Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G.Dorko, et al. (2005). The 2005 pascal visual object classes challenge. http ://www.pascal-network.org/challenges/VOC/.

Griffin, G., A. Holub, et P. Perona (2007). Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, http ://authors.library.caltech.edu/7694.

Hafiane, A., S. Chabrier, C. Rosenberger, et H. Laurent (2007). A new supervised evaluation criterion for region based segmentation methods. In *International conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*, pp. 439–448.

Hemery, B., H. Laurent, B. Emile, et C. Rosenberger (2009). Comparative study of local descriptors for measuring object taxonomy. In *IEEE International Conference on Image*

and Graphics (ICIG).

Hemery, B., H. Laurent, B. Emile, et C. Rosenberger (2010). Comparative study of localization metrics for the evaluation of image interpretation systems. *Journal of Electronic Imaging 19*(2), 023017.

Jurie, F. et C. Schmid (2004). Scale-invariant shape features for recognition of object categories. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) 2*, 90–96.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60*(2), 91–110.

Martin, D., C. Fowlkes, D. Tal, et J. Malik (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *International Conference on Computer Vision (ICCV) 2*, 416–423.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics 5*, 32.

Phillips, I. T. et A. K. Chhabra (1999). Empirical performance evaluation of graphics recognition systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 21*(9), 849–870.

Pratt, W., O. D. Faugeras, et A. Gagalowicz (1978). Visual discrimination of stochastic texture fields. *IEEE Transactions on Systems, Man, and Cybernetics (SMC) 8*(11), 796–804.

Riesen, K., M. Neuhaus, et H. Bunke (2007). Bipartite graph matching for computing the edit distance of graphs. *Lecture Notes in Computer Science : Graph-Based Representations in Pattern Recognition 4538*, 1–12.

Wilson, D. L., A. J. Baddeley, et R. A. Owens (1997). A new metric for grey-scale image comparison. *International Journal of Computer Vision 24*(1), 5–17.

Wolf, C. et J.-M. Jolion (2006). Object count/area graphs for the evaluation of object detection and segmentation algorithms. *International Journal on Document Analysis and Recognition 8*(4), 280–296.

Summary

Image processing algorithms include methods that process images from their acquisition to the extraction of useful information for a given application. Among these image interpretation algorithms, some are designed to detect, localize and identify one or several objects. The problem addressed here is the evaluation of interpretation results of an image or a video, given the ground truth. Challenges are multiple such as the comparison of algorithms, evaluation of an algorithm during its development or definition of its optimal setting. We propose a new metric for evaluating an interpretation result of an image. The advantage of the proposed metric is to evaluate a result by taking into account the quality of the localization, recognition and detection of objects of interest in the image. Several parameters allow us to change the behavior of this metric for a given application. Its behavior has been tested on a large database and is interesting.