Méthode d'analyse sémantique d'images combinant apprentissage profond et relations structurelles par appariement de graphes

Jérémy Chopin, Jean-Baptiste Fasquel, Harold Mouchère, Isabelle Bloch, Rozenn Dahyot

Méthode d'analyse sémantique d'images combinant apprentissage profond et relations structurelles par appariement de graphes

J. Chopin[1,2], J.-B. Fasquel[1], H. Mouchère[2], I. Bloch[3], R. Dahyot [4]

[1] LARIS, Université d'Angers, Angers, France
[2] LS2N, Université de Nantes, France
[3] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[4] School of Computer Science & Statistics, Trinity College Dublin, Ireland

Résumé

Nous proposons une méthode de segmentation sémantique d'images, combinant apprentissage profond et relations spatiales entre régions. Cette méthode repose sur l'appariement inexact de graphes, appliqué en sortie d'un réseau de neurones profond. Notre proposition est évaluée sur une base publique dédiée à la segmentation de visages, en mesurant l'IoU ("Intersection over Union") des boîtes englobantes des régions obtenues avec et sans utilisation des relations spatiales. Celles-ci permettent une amélioration de 2% en moyenne, et jusqu'à 24% dans certains cas.

Mots-clés

Vision par ordinateur, Apprentissage profond, Appariement inexact de graphes, Problème d'affectation quadratique.

Abstract

We propose a method for semantic image segmentation, combining a deep neural network and spatial relationships between image regions, encoded in a graph representation of the scene. Our proposal is based on inexact graph matching, applied to the output of a deep neural network. The proposed method is evaluated on a public dataset used for segmentation of images of faces. Preliminary results show that, in terms of IoU of region bounding boxes, the use of spatial relationships lead to an improvement of 2.4% in average, and up to 24.4% for some regions.

Keywords

Computer vision, Deep learning, Inexact graph matching, Quadratic assignment problem.

1 Introduction

L'apprentissage profond a montré son efficacité dans de nombreux domaines [8], en particulier pour la segmentation sémantique d'images en vision par ordinateur [7]. L'une des limites des approches par réseaux de neurones profonds est la nécessité de disposer d'un ensemble de données d'apprentissage vaste, représentatif et annoté [16]. De plus, la plupart de ces approches n'utilisent que les images elles-mêmes (les données), sans connaissance a priori sur les structures qu'elles contiennent et leur agencement spatial. De plus, ces méthodes nécessitent d'optimiser un grand nombre de paramètres.

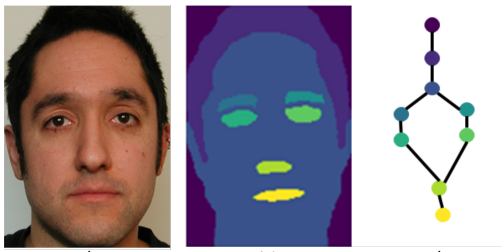

FIGURE 1 – Segmentation sémantique et relations spatiales. Les images sont extraites de la base de données publique FASSEG [10, 11]. a) Image initiale. b) Segmentation sémantique où chaque région appartient à une classe spécifique (par exemple l'oeil gauche, l'oeil droit). c) Relations spatiales modélisées par un graphe où chaque sommet correspond à une région spécifique de b : fond, cheveux, sourcil gauche/droit, oeil gauche/droit, nez, bouche. Les arêtes portent des relations correspondant aux distances entre les régions dans notre cas. Par souci de clarté, seules certaines arêtes du graphe complet sont affichées.

Ce type d'approche ignore les informations structurelles observables à haut-niveau. Cela peut concerner les relations spatiales entre différentes entités, comme l'illustre la figure 1 avec les positions relatives entre les principales régions du visage observées dans l'image annotée. Bien que souvent ignorées, ces informations structurelles ont montré leur potentiel dans de multiples travaux connexes, considérant des relations spatiales, d'inclusion ou de photométrie [1, 4, 5, 15], souvent appliquées à l'imagerie médicale [2, 3, 14]. Ces informations sont généralement représentées à l'aide de graphes, où les sommets correspondent à des régions et les arêtes portent les informations structurelles. L'identification des régions peut alors être réalisée par appariement de graphes [4, 5, 12], par raisonnement séquentiel dans les graphes, ou global par satisfaction de contraintes.

Dans ce contexte, nous proposons de combiner deux approches : les réseaux de neurones profonds qui s'avèrent

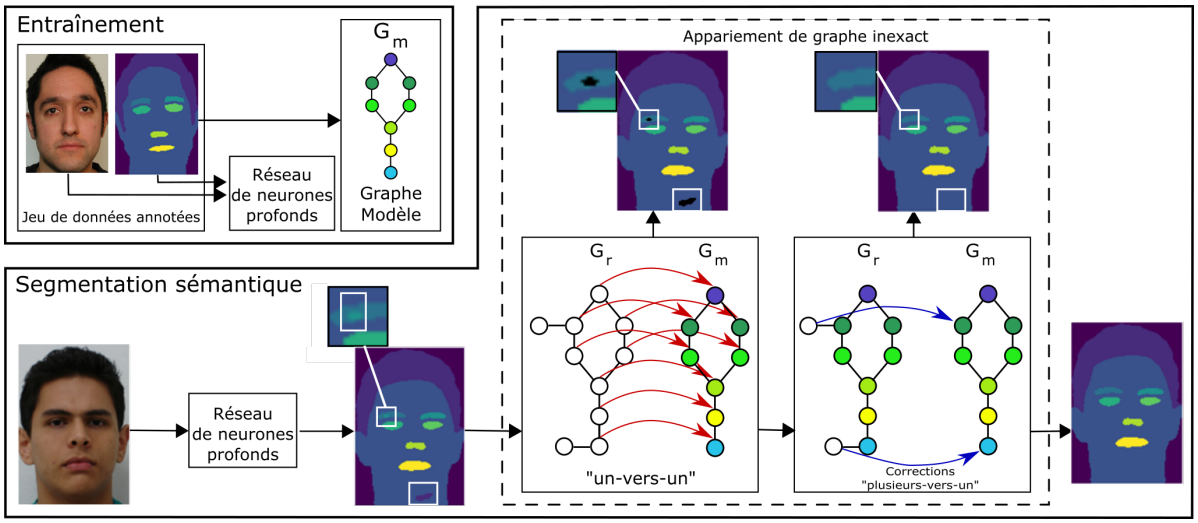

FIGURE 2 – Aperçu de la méthode. **Entraînement** : l'ensemble des données annotées est utilisé pour entraîner le réseau de neurones et construire un graphe modèle (similaire à celui de la figure 1-c). Par souci de clarté, bien que les graphes soient complets, seules quelques arêtes sont indiquées. La couleur des sommets correspond à la couleur des régions associées. **Segmentation sémantique** : le réseau de neurones produit une segmentation sémantique, éventuellement avec des artefacts (par exemple la région claire à l'intérieur du sourcil et la région sombre sur le cou). Un graphe G_r est ensuite construit à partir de cette segmentation et mis en correspondance avec le graphe modèle G_m. Cet appariement inexact entre graphes est réalisé en deux étapes : 1. les régions correctement segmentées sont récupérées (appariement un-vers-un), les artefacts étant ignorés (deux sommets restants dans cet exemple) ; 2. les artefacts restants sont mis en correspondance (appariement plusieurs-vers-un). Dans cet exemple, on peut voit que les deux artefacts sont correctement ré-étiquetés (voir les zones entourées).

efficaces mais nécessitent souvent de grands ensembles de données d'entraînement, et les graphes pour encoder des relations structurelles de haut niveau au sein d'images. Il est à noter que l'utilisation des réseaux de neurones profonds pour l'appariement de graphes est un sujet qui suscite actuellement beaucoup d'intérêt dans la communauté scientifique, pour des problèmes autres que la vision par ordinateur, par exemple en biologie, en sciences sociales, en linguistique [6, 9, 13]. A noter que ces travaux intègrent l'appariement de graphes dans le réseau de neurones profond, tandis que, dans notre cas, l'appariement est appliqué sur la sortie du réseau de neurones.

L'originalité de notre proposition concerne la formulation de la combinaison entre un réseau de neurones profond et l'appariement de graphes exploitant les relations spatiales, appliqué en sortie du réseau. Elle permet de corriger la segmentation sémantique, obtenue par l'utilisation de la carte de probabilité en sortie du réseau de neurones, en prenant en compte les relations spatiales observées dans la base de données annotées. L'utilisation de la structure spatiale globale de la scène permet également d'être moins sensible à la diversité (et donc à la taille) du jeu de données d'entraînement utilisé par le réseau de neurones.

La méthode proposée est détaillée dans la section 2. Des expériences préliminaires illustrant le potentiel de cette proposition sont présentées dans la section 3. La section 4 conclut le document par une discussion.

2 Apprentissage profond et connaissances structurelles

La figure 2 donne une vue d'ensemble de la méthode proposée avec les images considérées dans les expériences. À l'aide d'un ensemble de données d'apprentissage annoté, le réseau de neurones profond est entraîné à effectuer une segmentation sémantique. De plus, en utilisant uniquement les images annotées, les relations spatiales entre les différentes régions sont mesurées (telles que la moyenne des distances), ce qui conduit à un graphe modèle G_m où les sommets et les arêtes correspondent respectivement aux régions annotées et aux relations spatiales.

Lors du traitement d'une image inconnue, le réseau de neurones fournit une segmentation, à partir de laquelle un graphe d'hypothèse G_r est construit. Ce graphe est ensuite mis en correspondance avec le graphe modèle. L'objectif est de faire correspondre les sommets (et donc les régions sous-jacentes) produits par le réseau de neurones avec ceux du modèle, ce qui implique le ré-étiquetage de certaines régions (cf. appariement plusieurs-vers-un dans la figure 2). Cela produit une segmentation sémantique finale correspondant aux relations de haut niveau observées dans l'ensemble des données d'entraînement.

Nous détaillons ci-après l'étape de construction du graphe hypothèse à partir du réseau de neurones profond (Section 2.1) puis son appariement avec le graphe modèle (Section 2.2).

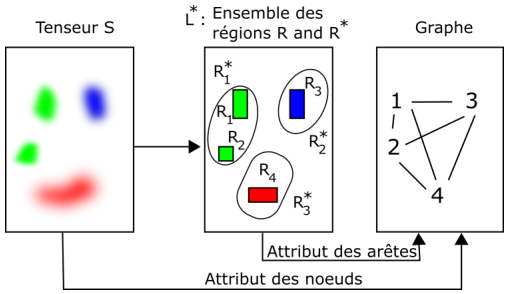

FIGURE 3 – Construction d'un graphe à partir du tenseur S et de la segmentation $\mathcal{L}^*$ qui en résulte. Chaque point de l'image de gauche est associé à un vecteur de probabilité représenté par des dégradés de couleurs. R_1^* est l'ensemble des régions (R_1 et R_2) qui appartiennent à la classe 1. Les attributs des arêtes sont calculés à partir des relations spatiales entre les régions R_i. Les attributs des sommets sont des vecteurs de probabilité moyens, calculés sur des régions R_i associées.

2.1 Construction des graphes

L'image d'entrée est traitée par le réseau de neurones qui produit, en sortie, un tenseur $S \in \mathbb{R}^{I \times J \times C}$ avec I la largeur (en pixels) de l'image, J la hauteur (en pixels) de l'image et C le nombre total de classes.
À l'emplacement du pixel (i, j), la valeur $S(i, j, c) \in [0, 1]$ est la probabilité d'appartenir à la classe c considérée dans la segmentation, avec les contraintes :

$$(\forall c = 1, \ldots, C,\ 0 \leq S(i, j, c) \leq 1) \wedge \left(\sum_{c=1}^{C} S(i, j, c) = 1 \right)$$

La carte de segmentation $\mathcal{L}^*$ sélectionne l'étiquette c de la classe ayant la plus forte probabilité :

$$\forall (i, j) \in \{1, ., I\} \times \{1, ., J\}, \quad \mathcal{L}^*(i, j) = \underset{c \in \{1, \ldots, C\}}{\arg\max}\, S(i, j, c)$$

À partir de cette carte de segmentation, nous définissons l'ensemble R de toutes les composantes connexes résultantes (voir figure 3, où $R = \{R_1, \cdots, R_4\}$). Nous définissons également un ensemble $R^* = \{R_1^*, \ldots, R_C^*\}$, où, pour chaque classe $c \in \{1, \cdots, C\}$, R_c^* est l'ensemble de régions correspondant aux composantes connexes appartenant à la classe c selon le réseau de neurones (voir figure 3, où $R^* = \{R_1^*, \ldots, R_3^*\}$). Cet ensemble R^* est utilisé pour contraindre l'appariement des graphes comme décrit dans la section 2.2.1.
À partir de l'ensemble R, une représentation structurelle est construite et modélisée par le graphe $G_r = (V_r, E_r, A, D)$, où V_r est l'ensemble des sommets, E_r l'ensemble des arêtes, A un interpréteur de sommets et D un interpréteur d'arêtes. Chaque sommet $v \in V_r$ est associé à une région $R_v \in R$ avec un attribut, fourni par la fonction A, qui est le vecteur de probabilité d'appartenance moyen sur l'ensemble des pixels $p = (i, j)$ composant R_v, donc calculé sur le tenseur S initial (voir figure 3) :

$$\forall v \in V_r, c \in \{1, \ldots, C\}, A(v)[c] = \frac{1}{|R_v|} \sum_{(i,j) \in R_v} S(i, j, c) \quad (1)$$

Nous considérons un graphe complet où chaque arête $e = (a, b) \in E_r$ a un attribut défini par la fonction D, associé à une relation entre les régions R_a et R_b (cf. figure 3). Dans notre cas, nous choisissons la distance minimale entre les deux régions :

$$\forall e = (a, b) \in E_r, D(e) = \min_{p \in R_a, q \in R_b} (\|p - q\|) \quad (2)$$

Le graphe modèle $G_m = (V_m, E_m, A, D)$, composé de C sommets (un sommet par classe), est construit à partir de l'ensemble d'entraînement. L'attribut d'un sommet sera un vecteur de dimension N avec une seule composante non nulle (de valeur égale à 1), associée à l'indice de la classe correspondante. Les arêtes E_m sont calculées en agrégeant les occurrences de relations entre les régions annotées des images d'entraînement. Dans cet article, la relation considérée est la distance minimale entre régions et l'agrégation est faite par leur moyenne.

2.2 Appariement avec le graphe modèle

Afin d'identifier les régions, l'objectif est d'associer chacun des sommets de G_r à un sommet du graphe modèle G_m. Selon l'hypothèse réaliste d'avoir plus de régions dans l'image associée à G_r que dans le modèle (c'est-à-dire $|V_r| \geq |V_m|$), nous sommes confrontés à un problème d'appariement inexact de graphes de type plusieurs-vers-un [12]. Nous proposons de formuler cet appariement comme un problème d'affectation quadratique (QAP, pour *Quadratic Assignement Problem*), comme cela a été récemment considéré [17].
Dans notre cas, la mise en correspondance est représentée par une matrice $X \in \{0, 1\}^{|V_r| \times |V_m|}$, où $X_{ij} = 1$ signifie que le sommet $i \in V_r$ est mis en correspondance avec le sommet $j \in V_m$. Cela est illustré dans la figure 4 dans deux cas (correspondances "un-vers-un" et "plusieurs-vers-un"). L'objectif est de déterminer la meilleure correspondance (X^*), solution de :

$$X^* = \arg\min_X \left\{ \mathrm{vec}(X)^T K\, \mathrm{vec}(X) \right\} \quad (3)$$

où $\mathrm{vec}(X)$ est la représentation sous forme de vecteur colonne de la matrice X et T est l'opérateur de transposition. La matrice K, non détaillée ici par souci de concision (voir [17] pour plus de détails), intègre les mesures de dissimilarité entre les deux graphes G_r et G_m, au niveau des sommets (éléments diagonaux) et des arêtes (éléments non diagonaux) :

$$K = \alpha\, K_v + (1 - \alpha)\, \frac{K_e}{\max K_e} \quad (4)$$

où K_v intègre les dissimilarités entre les sommets (distance euclidienne entre les vecteurs de probabilité d'appartenance à une classe), et $\alpha \in [0, 1]$ un paramètre. Dans

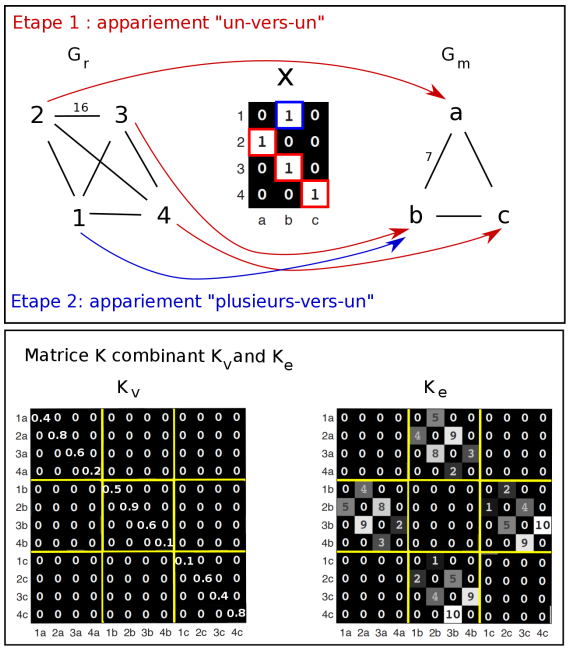

FIGURE 4 – Appariement de graphes formulé comme un QAP (illustration inspirée de [17]), où la matrice X matérialise un appariement entre les graphes G_r et G_m. La première étape (en rouge) vise à chercher une correspondance de type "un-vers-un" (chaque sommet de V_m est associé à un seul sommet de V_r). La deuxième étape (en bleu) vise à faire correspondre les sommets restants de G_r, ce qui conduit à un appariement final de type "plusieurs-vers-un". Pour trouver l'appariement optimal, la matrice K est utilisée, combinant les matrices K_v et K_e, qui mesurent respectivement les dissimilarités entre les sommets et les arêtes. Par souci de clarté, seuls les attributs scalaires de deux arêtes sont indiqués.

l'exemple considéré dans la figure 4, $K_v[1,1] = 0,4$ (ligne et colonne nommées `2a`) représente la dissimilarité, en termes de vecteurs de probabilité, entre les sommets 2 de G_r et a de G_m, si l'on faisait correspondre ces deux sommets.

La matrice K_e est liée aux dissimilarités entre les arêtes. Par exemple, dans la figure 4, $K_e[6,1] = 9$ (ligne et colonne respectivement nommées `3b` et `2a`) correspond à la dissimilarité entre les arêtes $(2,3) \in E_r$ (attribut scalaire valant 16) et $(a,b) \in E_m$ (attribut scalaire valant 7), si nous faisions correspondre simultanément le sommet 2 avec le sommet a et le sommet 3 avec le sommet b. Dans un tel cas, $K_e[6,1]$ est calculé en utilisant ces deux attributs : $K_e[6,1] = 16 - 7 = 9$. Les termes K_e sont liés aux distances entre les régions (normalisées dans la matrice finale K).

Le paramètre α ($\alpha \in [0,1]$) permet de pondérer la contribution relative des dissimilarités entre les sommets et les arêtes (les termes K_v varient entre 0 et 1, et K_e est normalisé selon l'équation 4).

En raison de la nature combinatoire de ce problème d'optimisation [17], nous proposons une procédure en deux étapes, s'appuyant sur la segmentation sémantique initiale fournie par le réseau de neurones, et consistant à :

1. rechercher une première correspondance "un-vers-un" (cf. figure 4- étape 1) ;
2. affiner l'appariement par la mise correspondance des sommets restants, conduisant finalement à une correspondance "plusieurs-vers-un" (cf. figure 4- étape 2).

2.2.1 Appariement initial : "un-vers-un"

On cherche la solution optimale à l'équation 3 en imposant les trois contraintes suivantes à X, réduisant ainsi l'espace de recherche pour les candidats éligibles :

1. $\sum_{j=1}^{|V_m|} X_{ij} \leq 1$: certains sommets i de G_r peuvent n'être associés à aucun sommet de G_m.
2. $\sum_{i=1}^{|V_r|} X_{ij} = 1$: chaque sommet j de G_m doit être associé à un seul sommet de G_r (cas des 1 entourés en rouge dans la figure 4).
3. $X_{ij} = 1 \Rightarrow R_i \in R_j^*$: le sommet $i \in V_r$ peut être mis en correspondance avec le sommet $j \in V_m$ si la région R_i associée a été initialement considérée par le réseau de neurones comme appartenant très probablement à la classe j ($R_i \in R_j^*$). Par exemple, dans le cas de la figure 3, seuls les sommets liés aux régions R_1 et R_2 seraient considérés comme candidats pour la classe 1 (R_1^*).

Les deux premières contraintes 1 et 2 assurent la recherche d'un appariement de type "un-vers-un". Grâce à la troisième contrainte, on réduit l'espace de recherche en s'appuyant sur le réseau de neurones : on suppose qu'il a correctement, au moins dans une certaine mesure, identifié les régions cibles, même si des artefacts peuvent encore avoir été produits (à gérer en affinant ultérieurement l'appariement). Cette étape nous permet de retrouver la structure générale des régions (modélisée par G_m).

2.2.2 Appariement final : "plusieurs-vers-un"

Nous associons chaque sommet k restant ($k \in V_r \mid \sum_{j=1}^{|V_m|} X_{kj} = 0$) au sommet $i^* \in V_r$ en considérant la fonction de coût suivante entre deux sommets i et k de G_r :

$$\text{cost}(i,k) = \alpha |A(k) - A(i)| + (1-\alpha) \frac{D(k) - D(i)}{\max_{u \in E_r} D(u)}. \quad (5)$$

$$i^* = \underset{i \in V_r \mid \sum_{k=1}^{|V_m|} X_{ik} = 1}{\text{argmin}} \text{cost}(i,k). \quad (6)$$

Selon cette formulation, il apparaît que les sommets restants sont appariés aux sommets de G_m en recherchant indirectement des correspondances avec les sommets déjà appariés de G_r. Par conséquent, on se concentre sur les similarités au sein de l'image actuelle et non avec le modèle. Dans la figure 4, cela correspond à trouver la correspondance entre les sommets 1 et b en étudiant indirectement la pertinence de la correspondance des sommets 1 et 3 (le sommet 3 étant déjà apparié avec le sommet b).

La formulation de l'équation 6 et concernant un seul sommet est similaire à la formulation matricielle des équations 3 et 4 (et concernant simultanément plusieurs sommets). La seule différence est que nous considérons les dissimilarités des sommets et des arêtes dans le graphe G_r au lieu de considérer les dissimilarités entre G_r et G_m.

3 Expérimentations

Nous présentons, ci-après, l'ensemble de données considéré, puis le protocole d'évaluation, et enfin les résultats.

3.1 Données

Nous considérons le jeu de données public FASSEG [1]. Celui-ci porte sur la segmentation sémantique multi-classes du visage [10] (cf. figure 1) ainsi que l'estimation de sa pose [11]. Pour cette étude préliminaire, nous considérons un sous-ensemble de ce jeu de données qui correspond à une pose spécifique (la vue de face) et contient 70 images.

FASSEG ne permet cependant pas de distinguer certaines régions du visage (c'est-à-dire l'œil gauche et l'œil droit, le sourcil gauche et le sourcil droit). Nous avons donc affiné les annotations afin de donner un label unique à ces régions (voir figure 5).

Image Initial labels Modified labels

FIGURE 5 – Étiquettes FASSEG modifiées.

3.2 Protocole d'évaluation

Pour les expériences, nous considérons le réseau neuronal U-Net [16] qui s'adapte bien à un ensemble d'entraînement avec un petit nombre d'échantillons. Nous avons aussi divisé notre ensemble de données comme suit : 20 images sont utilisées pour l'ensemble d'entraînement de référence, 10 pour l'ensemble de validation de référence et 40 pour l'ensemble de test de référence. De plus, 100 époques sont utilisées pour la formation du réseau.

Le graphe modèle est construit en calculant les distances moyennes (équation 2) entre les différentes régions annotées de l'ensemble d'entraînement. Le paramètre α est choisi empiriquement, sur la base d'observations sur certaines images, et est fixé à $0,4$ pour les expériences, mais sans aucune optimisation.

Comme cela a été étudié expérimentalement, une simple mesure de la distance entre les barycentres (centres de gravité) des régions est apparue inappropriée, en raison de la variabilité de la forme de certaines régions (par exemple les cheveux).

Nous évaluons la différence entre la qualité de la segmentation sémantique à la sortie des réseaux de neurones et celle après la mise en correspondance, c'est-à-dire avec l'intégration de l'information structurelle. Nous considérons l'*Intersection over Union* (IoU ou indice de Jaccard) pour évaluer la qualité de nos résultats par rapport à notre annotation manuelle utilisée comme vérité terrain. Cette mesure d'évaluation est utilisée pour comparer les régions au niveau des pixels et également au niveau des boîtes englobantes. Cette comparaison des boîtes englobantes nous permet de quantifier les erreurs de segmentation correspondant à une région principale correcte mais avec des erreurs liées à une ou plusieurs sous-régions éloignées de la région principale, et comprenant peu de pixels. Ces mesures sont effectuées pour chaque classe, la valeur moyenne globale étant également calculée.

Nous étudions également l'impact de la taille de la base d'apprentissage sur la qualité de notre segmentation sémantique. Des expériences sont réalisées pour différentes tailles, exprimées en pourcentage de la base de référence, à savoir 100 % (totalité des 20 images de l'ensemble d'entraînement et totalité des 10 images pour la validation), 75 % (15 images de l'ensemble d'entraînement et 7 images de l'ensemble de validation), 50 % et 25 % (5 images de l'ensemble d'entraînement et 2 pour la validation). Pour les tailles allant de 75 % à 25 %, les expériences sont effectuées 20 fois avec un tirage aléatoire des images utilisées pour l'entraînement et la validation, et les performances moyennes sont retenues.

3.3 Résultats quantitatifs

Les tables 1 et 2 fournissent des résultats respectivement en termes d'indice IoU pixel par pixel et d'indice IoU en termes de boîtes englobantes. Dans l'ensemble, la segmentation sémantique a été améliorée grâce à l'appariement de graphes, comme on peut le constater en moyenne sur l'ensemble des classes.

L'amélioration apparaît plus significative lorsque l'on utilise l'IoU sur les boîtes englobantes des régions (table 2) par rapport à l'IoU sur les régions elles-mêmes (table 1). Cela est dû au fait que, pour une classe donnée, l'IoU sur les boîtes englobantes est très sensible à la répartition spatiale des zones mal classées, même si ces zones sont de petite taille par rapport à l'ensemble de la région. La mesure d'IoU sur les régions est beaucoup moins sensible à ces erreurs, en particulier pour ces zones d'erreur de petite taille (d'autant plus pour des classes de grande taille telles que les cheveux ou le visage, contrairement aux sourcils et aux yeux par exemple). Ainsi, on peut noter, en considérant les boîtes englobantes, une amélioration très significative de $24,4$ % pour la classe *sourcil droit* (R-Br) avec une taille de jeu d'apprentissage de 50 % (table 2). L'utilisation de relations spatiales permet d'éviter les régions mal classées et spatialement incohérentes (telles que des cheveux trouvés entre le nez et l'oeil droit).

Les deux tables 1 et 2 illustrent également l'influence de la

1. L'ensemble de données publiques annotées FASSEG peut être téléchargé à l'adresse suivante : https ://github.com/massimomauro/FASSEG-dataset.

TABLE 1 – Comparaison, mesurée à l'aide de l'IoU sur les régions, entre notre segmentation manuelle (de référence) et celle fournie par le U-Net uniquement (notée `U-Net`) ainsi que celle résultant de notre approche (notée `U-Net+GM`, où GM signifie *graph matching*). Les résultats sont fournis sous forme de moyenne et pour chaque classe : Bg (fond), Hr (cheveux), Fc (visage), L-br (sourcil gauche), R-br (sourcil droit), L-eye (oeil gauche), R-eye (oeil droit), Nose (nez) et Mouth (bouche). Les résultats sont également fournis pour les différentes tailles de base d'apprentissage.

		Classes									
Ensemble d'apprentissage (%)	Approche	Moyenne	Bg	Hr	Fc	L-br	R-br	L-eye	R-eye	Nose	Mouth
100	U-Net	75,3	88,0	88,2	**91,9**	61,8	60,5	76,9	72,8	67,3	77,2
	U-Net + GM	**75,4**	88,0	**88,9**	91,8	**62,1**	**60,7**	**77,0**	72,8	67,3	77,2
75	U-Net	74,0	**88,5**	85,9	91,0	60,8	56,8	75,5	72,8	64,7	77,9
	U-Net + GM	**74,3**	88,3	**86,8**	**91,1**	**61,7**	**57,5**	**75,6**	72,8	64,7	**78,0**
50	U-Net	72,0	86,0	84,9	90,9	54,6	54,2	73,5	72,4	65,9	75,4
	U-Net + GM	**73,7**	**86,7**	**86,9**	**91,0**	**59,9**	**57,9**	**74,5**	**72,5**	**66,1**	**75,6**
25	U-Net	38,2	**84,5**	83,6	56,8	2,4	28,1	61,7	25,6	**57,7**	70,0
	U-Net + GM	**39,8**	83,8	**86,4**	**61,0**	**2,7**	**30,6**	**65,7**	**26,5**	55,1	**71,4**

TABLE 2 – Comparaison, mesurée à l'aide de l'indice IoU sur les boîtes englobantes des régions, entre la segmentation manuelle et celle fournie par le U-Net (`U-Net`) ainsi que celle résultant de notre approche (`U-Net+GM`, où GM signifie *graph matching*). Les résultats sont fournis sous forme de moyenne et pour chaque classe : Bg (fond), Hr (cheveux), Fc (visage), L-br (sourcil gauche), R-br (sourcil droit), L-eye (oeil gauche), R-eye (oeil droit), Nose (nez) et Mouth (bouche). Les résultats sont également fournis pour les différentes tailles de base d'apprentissage.

		Classes									
Ensemble d'apprentissage (%)	Approche	Moyenne	Bg	Hr	Fc	L-br	R-br	L-eye	R-eye	Nose	Mouth
100	U-Net	76,0	84,0	82,5	96,1	66,3	63,6	74,0	75,1	69,0	78,2
	U-Net + GM	**78,4**	84,0	**92,4**	96,1	**66,5**	**68,3**	**79,4**	**75,2**	**70,4**	**78,9**
75	U-Net	74,7	84,0	74,5	95,8	66,7	59,5	78,0	**75,1**	64,9	79,8
	U-Net + GM	**77,5**	84,0	**84,9**	**96,1**	**66,8**	**69,2**	**78,8**	75,0	**67,5**	**79,9**
50	U-Net	68,0	85,4	75,7	**94,5**	52,4	41,8	77,7	66,6	63,8	70,6
	U-Net + GM	**76,5**	85,4	**86,0**	94,4	**65,3**	**66,2**	**78,5**	**74,9**	**68,5**	**74,7**
25	U-Net	36,3	**81,2**	80,1	**90,5**	1,7	33,7	**68,4**	12,8	55,1	66,9
	U-Net + GM	**42,3**	78,0	**90,1**	82,8	**2,9**	**34,8**	65,9	**27,4**	**59,3**	**70,1**

réduction de la base d'apprentissage. Il apparaît que plus la base est réduite, plus l'apport des relations spatiales de haut niveau est important (cela étant très significatif dans la table 2).
Il est à noter que, pour une base d'apprentissage de seulement 25 %, en fonction d'un tirage aléatoire réalisé, certains réseaux semblaient incapables de proposer des régions candidates pour certaines classes sur certaines images de test. Cela est dû à la faible représentativité de la base d'apprentissage. Dans ces cas, la mise en correspondance des graphes échoue en raison de l'absence de candidats pour certaines classes, ce qui n'est pas encore géré par notre approche. C'est pourquoi, pour un taux de 25 %, les résultats présentés dans les tables 1 et 2 ignorent ces cas, et les performances rapportées ne sont moyennées que sur des images de test segmentées fournissant au moins une région par classe (48,9 % des images ne sont pas prises en compte).

3.4 Résultats qualitatifs de la segmentation

Les figures 6 et 7 donnent quelques exemples de segmentations sémantiques avec le U-Net et notre approche.
La figure 6 donne quelques exemples représentatifs. Comme on peut l'observer visuellement, les améliorations sont significatives dans de nombreux cas (cf. figure 6 - image 75 % : cheveux en bas du visage). Même pour un petit ensemble de données d'entraînement, les informations structurelles peuvent améliorer la segmentation, parfois seulement partiellement dans le cas de sorties U-Net vraiment dégradées (cf. figure 6 - image 25 %).

La figure 7 se concentre sur la réduction de la base de données d'apprentissage pour illustrer la robustesse de notre approche. Ces exemples représentatifs illustrent comment la réduction de la base de données détériore la segmentation en utilisant uniquement des réseaux neuronaux profonds. Cela montre comment notre approche, intégrant des contraintes spatiales globales, peut compenser de manière significative la perte de diversité de l'ensemble des données d'apprentissage.

3.5 Complexité algorithmique

La table 3 indique les durées d'exécution mesurées (moyenne sur les 40 images utilisées pour les tests), pour différentes tailles de la base de données d'apprentissage. Il s'agit d'une estimation brute, pour ce domaine d'application, des durées d'exécution sans détailler le coût de chaque étape, et sans aucune tentative d'optimisation du code, écrit en Python (utilisation de `numpy` pour l'algèbre linéaire). A noter que les programmes ont été exécutés sur un noyau Intel I7, sans GPU. Les durées d'exécution sont plus faibles pour une grande taille de base de données d'apprentissage, car le réseau U-Net produit moins d'artefacts, ce qui réduit le nombre de sommets dans G_r et donc le temps d'appariement.

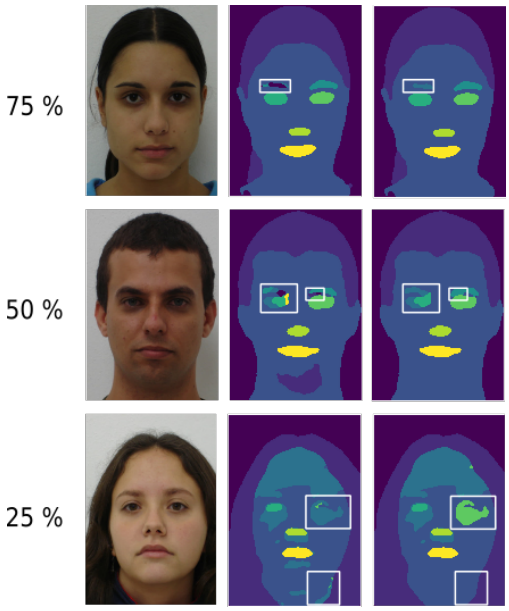

FIGURE 6 – Exemples de résultats de segmentation (image initiale, résultat avec U-Net, résultat avec l'approche proposée), pour différentes tailles de la base d'apprentissage (75%, 50% et 25%).

TABLE 3 – Durée d'exécution de l'ensemble des données de test, en secondes. Moins nous utilisons de données pendant la phase d'apprentissage, plus il faudra de temps pour traiter une image de test, comme le montre l'augmentation du temps moyen.

Apprentissage (%)	Moyenne	Ecart-type	Médiane	Min	Max
100	28,5	2,3	27,8	25,4	34,1
75	39,4	29,8	27,4	25,1	372,6
50	81,0	791,0	31,0	24,7	20165,9
25	171,5	599,2	49,1	25,3	7363,8

4 Discussion et perspectives

L'approche proposée semble efficace pour améliorer la segmentation sémantique réalisée par le réseau neuronal profond U-Net. L'amélioration semble particulièrement significative pour certaines classes (jusqu'à $24,4$ % pour la classe *sourcil droit*, comparé à l'amélioration moyenne de 8,5 % en considérant l'IoU sur les boîtes englobantes avec un pourcentage de la base d'apprentissage de 50 %). Cela est dû à la nature des informations considérées, qui semblent très complémentaires aux informations de bas niveau (au niveau des pixels) considérées par le U-Net, qui ignore intrinsèquement les relations spatiales observées à haut niveau. Un autre point fort de notre approche est sa capacité à compenser les erreurs de segmentation résultant de la réduction de la base de données d'apprentissage. C'est un aspect important car sa taille est une limitation forte de l'apprentissage profond, comme cela a été souligné dans l'introduction.

Bien que prometteuse, notre approche présente quelques limites. Premièrement, notre approche est invariante en translation et en rotation, mais pas encore en échelle. On pourrait y remédier en introduisant un facteur d'échelle (à estimer automatiquement) dans le processus d'appariement (notamment en ce qui concerne le calcul de la matrice K figurant dans l'équation 3).

Deuxièmement, les occultations partielles dans les images qui pourraient affecter le calcul du graphe G_r sont des cas qui ne sont pas encore pris en compte dans notre formulation actuelle. En effet, on suppose que chaque sommet (région) correspond nécessairement à une région du modèle. Le traitement des occultations pourrait être géré en assouplissant les hypothèses relatives à l'appariement aux sommets du modèle.

Troisièmement, nous n'avons pas étudié, dans cet exemple applicatif, la robustesse de notre approche vis-à-vis du changement de pose qui pourrait modifier les relations spatiales. Il est à noter que d'autres domaines d'application de notre technique, comme la segmentation d'images médicales en 3D, peuvent ne pas présenter les mêmes défis en pratique que la segmentation de visage présentée ici. Pour supporter une forte variation dans les relations spatiales, une solution pourrait être de considérer un ensemble de graphes modèles représentatifs au lieu d'un seul, avec la difficulté sous-jacente de choisir le modèle approprié lors de la segmentation d'une image.

Enfin, la réduction du temps de calcul peut être cruciale dans certains scénarios d'application. En particulier, la grande complexité de la première étape (formulée comme un QAP) peut impliquer une augmentation considérable du temps de calcul, si le nombre de classes et de régions augmente. Néanmoins, les estimations brutes fournies dans le document sont encourageantes et des améliorations supplémentaires pourraient être recherchées en utilisant du matériel spécialisé (par exemple, GPU).

5 Conclusion

Dans cet article, nous avons proposé une méthode originale combinant les relations spatiales avec un apprentissage profond pour la segmentation sémantique. Les résultats préliminaires montrent le potentiel de cette approche, avec des améliorations significatives dans certains cas tout en nous permettant de réduire la taille de la base d'apprentissage requise.

L'amélioration obtenue est de 8,5% en moyenne, par rapport à la segmentation sémantique résultant du réseau neuronal U-Net, entraîné avec seulement 50% de l'ensemble de données d'apprentissage disponible.

Les travaux futurs se concentreront sur l'invariance par changement d'échelle, une évaluation plus fine de la complexité et sur des expériences dans d'autres domaines d'application.

Remerciements

Ces travaux ont été menés dans le cadre du programme région Atlanstic2020 (Recherche, Formation et Innovation en Pays de la Loire), soutenu par la région des Pays de la Loire et le fonds européen de développement.

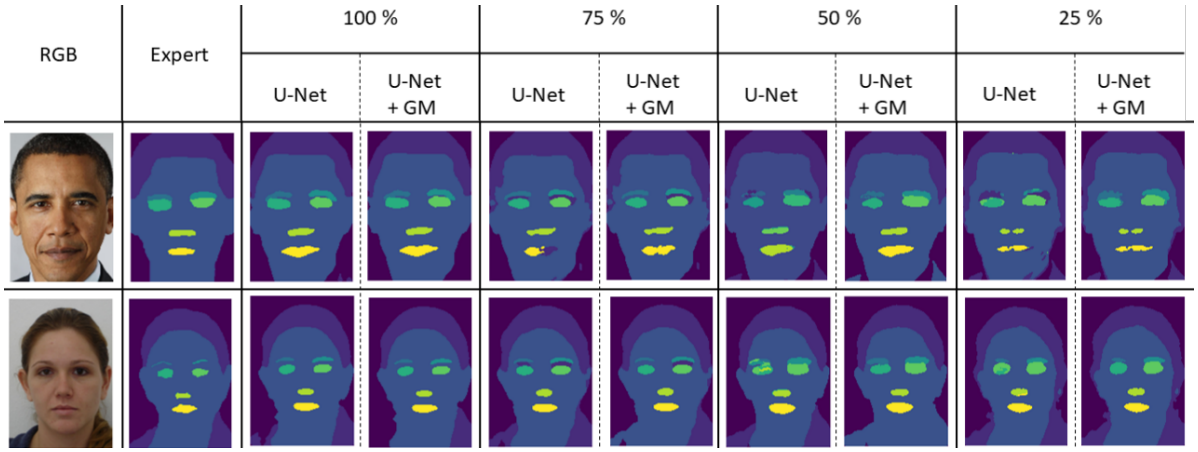

FIGURE 7 – Influence de la taille de la base de données d'apprentissage (de 100 % à 25 %) avec comme exemples des résultats obtenus sur 2 images représentatives.

Références

[1] I. Bloch. Fuzzy sets for image processing and understanding. *Fuzzy Sets and Systems*, 281 :280–291, 2015.

[2] O. Colliot, O. Camara, and I. Bloch. Integration of fuzzy spatial relations in deformable models - application to brain MRI segmentation. *Pattern Recognition*, 39 :1401–1414, 2006.

[3] J.-B. Fasquel, V. Agnus, J. Moreau, L. Soler, and J. Marescaux. An interactive medical image segmentation system based on the optimal management of regions of interest using topological medical knowledge. *Computer Methods and Programs in Biomedicine*, 82 :216–230, 2006.

[4] J.-B. Fasquel and N. Delanoue. An approach for sequential image interpretation using a priori binary perceptual topological and photometric knowledge and k-means based segmentation. *Journal of the Optical Society of America A*, 35(6) :936–945, 2018.

[5] J.-B. Fasquel and N. Delanoue. A graph based image interpretation method using a priori qualitative inclusion and photometric relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5) :1043–1055, 2019.

[6] H. Gao and S. Ji. Graph U-Nets. In K. Chaudhuri and R. Salakhutdinov, editors, *36th International Conference on Machine Learning*, volume 97, pages 2083–2092, 2019.

[7] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70 :41 – 65, 2018.

[8] I. J. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016.

[9] P. Goyal and E. Ferrara. Graph embedding techniques, applications, and performance : A survey. *Knowledge-Based Systems*, 151 :78 – 94, 2018.

[10] K. Khan, M. Mauro, and R. Leonardi. Multi-class semantic segmentation of faces. In *IEEE International Conference on Image Processing*, pages 827–831, 2015.

[11] K. Khan, M. Mauro, P. Migliorati, and R. Leonardi. Head pose estimation through multi-class face segmentation. In *IEEE International Conference on Multimedia and Expo*, pages 175–180, 2017.

[12] O. Lezoray and L. Grady. *Image Processing and Analysis with Graphs : Theory and Practice*. Digital Imaging and Computer Vision. CRC Press, 2012.

[13] Z. Li, L. Zhang, and G. Song. GCN-LASE : Towards adequately incorporating link attributes in graph convolutional networks. In *Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2959–2965, 2019.

[14] A. Moreno, C.M. Takemura, O. Colliot, O. Camara, and I. Bloch. Using anatomical knowledge expressed as fuzzy constraints to segment the heart in CT images. *Pattern Recognition*, 41(8) :2525 – 2540, 2008.

[15] O. Nempont, J. Atif, and I. Bloch. A constraint propagation approach to structural model based image segmentation and recognition. *Information Sciences*, 246 :1–27, 2013.

[16] O. Ronneberger, P. Fischer, and T. Brox. U-Net : Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[17] F. Zhou and F. De la Torre. Factorized graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9) :1774–1789, 2016.