

Vision Gestalt et connaissances : une approche générique à l'interprétation d'images

N. Zlatoff B. Tellez A. Baskurt

LIRIS, Université Claude Bernard, Lyon 1
Bât. Nautibus, 8 boulevard Niels Bohr, 69622 Villeurbanne cedex – France

{nzlatoff, btellez, abaskurt}@liris.cnrs.fr

Résumé

Le fossé sémantique existant entre les descripteurs bas-niveaux et les concepts sémantiques rend ambigu la mise en correspondance des régions d'une image segmentée avec un objet sémantique. Aussi, l'introduction de connaissances du domaine dans le processus semble inévitable. Toutefois, de tels systèmes sont souvent dépendants d'un domaine particulier, car les connaissances sont directement incluses dans les procédures de traitement. Nous proposons donc un système générique basé sur la théorie de la vision Gestalt, capable d'évaluer la pertinence de ses groupements par la connaissance du domaine. Afin de pouvoir être utilisée de manière transparente et flexible, cette connaissance est externalisée dans des ontologies. En outre, nous insistons sur la nécessité de modéliser cette connaissance comme connaissance de scène sur une hiérarchie multi-niveaux. Nous présentons l'architecture générale de notre système ainsi qu'un exemple d'application dans le domaine de l'indexation d'image de stèles thessaliennes.

Mots clefs

Interprétation d'images, sémantique, vision Gestalt, connaissances.

1 Introduction

Durant ces dix dernières années, les augmentations conjointes des capacités de stockage et de l'interconnexion des réseaux ont rendu disponible un volume considérable de données numériques, et tout spécialement d'images.

Des systèmes d'indexation, permettant à des utilisateurs de trouver les images pertinentes en réponse à une requête, sont ainsi fortement demandés. Aujourd'hui, de nombreux outils basés sur le contenu existent. Ils décrivent une image sur des critères bas-niveaux comme le couleur, la texture voire la forme d'un objet principal [1]. Toutefois, de tels systèmes ne sont pas totalement pertinents car un utilisateur cherche généralement une image d'après ce qu'elle représente (sa sémantique) et non par son aspect de couleur ou de texture.

Chercher à étendre le principe des outils basés sur le contenu pour dériver de la sémantique directement à partir de critères bas-niveaux est une technique qui trouve rapidement ses limites. En effet, le fossé sémantique stipule qu'un tel lien n'existe pas sans ambiguïté.

Ainsi, il est indispensable d'introduire des connaissances relatives au domaine d'application afin de déduire ce que représente une partie d'image, dans un contexte donné.

1.1 Etat de l'art

Les années 80 ont vu l'apparition de nombreux outils vision basés sur les connaissances. Ainsi, SIGMA [2] ou Schema [3] sont des systèmes d'interprétation d'images aériennes. Ils se basent sur des descriptions *ad hoc* de classes d'objets susceptibles d'apparaître dans les images, auxquelles sont attachées des caractéristiques communes aux objets de la classe et des procédures de contrôle, précisant quelles caractéristiques rechercher et dans quel ordre.

Néanmoins, [4] remarque que ces systèmes sont fortement dépendants du domaine traités, puisqu'ils incorporent des connaissances *a priori* sur la scène ou les objets du domaine, directement dans les procédures de traitement.

Dans cette optique, des travaux sur des algorithmes génériques sont apparus : certains d'entre eux tentent de grouper des régions issues d'une image segmentée, sans aucune considération du domaine, en se basant sur la maximisation de la probabilité des groupements [5]. La plupart du temps toutefois, les travaux font appel, de près ou de loin, à la théorie de la vision Gestalt [6], selon laquelle la vision humaine crée des groupements successifs (*gestalts*). Ces derniers sont basés sur cinq propriétés : proximité, similarité, fermeture, continuité et symétrie (voir Figure 1). Ainsi, chaque objet acquiert une pertinence vis-à-vis d'un contexte global. Dans cette optique, [7] et [8] présentent un système de regroupement Gestalt sur des segments. Dans [7], la pertinence des groupements est évaluée par la théorie de Dempster-Schafer, alors que [8] contrôle le processus avec des MRF (Markov Random Fields). Dans les deux cas, les mécanismes de groupement sont totalement déconnectés des connaissances relatives aux objets qu'ils manipulent.

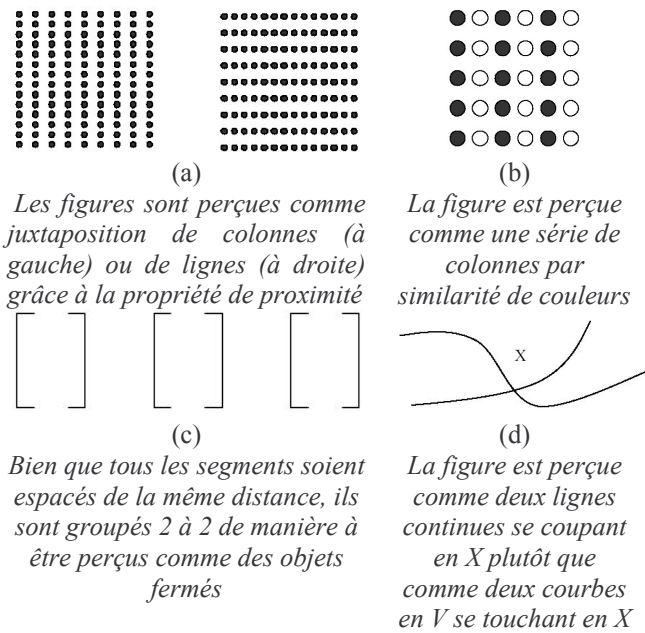


Figure 1- Illustration des propriétés Gestalt de regroupement : proximité (a), similarité (b), fermeture (c) et continuité (d).

1.2 Architecture générique

Au contraire, nous pensons que les systèmes d'interprétation d'images *connaissent* ce qu'ils perçoivent. Ainsi, nous proposons un processus générique de groupement de régions (issues d'une image segmentée), sous le contrôle des connaissances du domaine dénoté par l'image en cours de traitement. Ainsi, le système est capable de s'interfacer à différentes connaissances domaine, en fonction de l'image traitée, et d'en extraire toute information nécessaire à son fonctionnement (Figure 2).

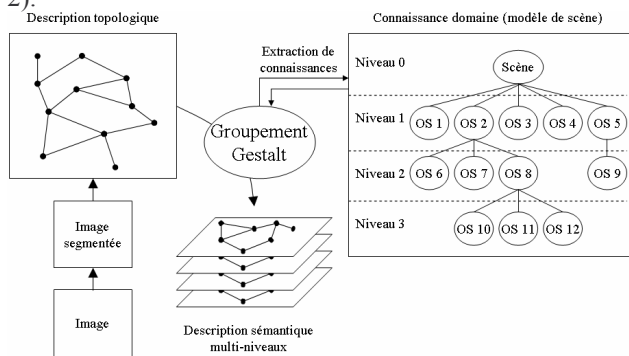


Figure 2- Architecture générique

Les connaissances du domaine sont modélisées dans une ontologie et doivent être mises en correspondance avec la description issue de l'image segmentée (c'est-à-dire un graphe d'adjacence avec des descripteurs bas-niveaux additionnels). C'est pourquoi la connaissance du domaine doit être modélisé comme une connaissance de scène qui précise comment les objets apparaissent dans les images.

En outre, il apparaît que tous les objets ne sont pas pertinents à tous les niveaux de détails et ne doivent donc pas tous être perçus à tous les niveaux : certains sont des composants principaux, d'autres des éléments de détails. Dans cette optique, nous modélisons la notion de composition de la scène, en organisant les objets sémantiques sur différents niveaux. Les processus de groupement Gestalt seront ensuite effectués à chacun des niveaux de la hiérarchie, depuis le plus général. Cette approche permet de fournir un contexte d'interprétation plus fin à chaque étape et aboutit également à une description sémantique sur plusieurs niveaux.

2 Mécanisme générique de groupement

Un tel algorithme autorise le groupement itératif de régions issues d'une segmentation. Nous proposons un mécanisme de groupement basé sur la théorie Gestalt sous le contrôle des connaissances du domaine.

La théorie Gestalt a été introduite par Wertheimer [6] au début du vingtième siècle. Elle considère que le mécanisme de vision crée des regroupements successifs (*gestalts*) de stimuli, selon cinq propriétés : proximité, similarité, fermeture, continuité et symétrie. Toutefois, ces propriétés correspondent à des notions relativement haut-niveau, et leur implantation ne va pas de soi.

Toujours selon la théorie Gestalt, la proximité est l'une des propriétés les plus importantes. C'est pourquoi nous choisissons dans un premier temps de travailler sur un graphe d'adjacence (RAG) et de le réduire itérativement en regroupant les régions d'après les autres propriétés Gestalt. Nous avons implanté à ce stade les propriétés de similarité, fermeture et continuité dans une version simplifiée. Ces implantations devront être plus détaillées dans un deuxième temps.

Plus précisément, l'algorithme de groupement (inspiré de [9] et étendu dans une perspective gestalt) calcule pour chaque frontière entre deux régions i et j du RAG, une distance Gestalt qui prend en compte les différentes propriétés gestalts :

$$DG_{ij} = S_{ij} \times CL_{ij} \times CO_{ij}$$

$$S_{ij} = \sqrt{\sum_{k=0}^n (d_{j,k} - d_{i,k})^2} \text{ avec } d_{j,k} \text{ le } k^{\text{ième}} \text{ descripteur de la}$$

région j et n le nombre total de descripteurs. S_{ij} est une distance euclidienne permettant de mesurer la similarité entre deux régions.

$$CL_{ij} = \frac{\min(P_i, P_j)}{4P_{ij}} \text{ avec } P_i \text{ le périmètre de la région } i \text{ et } P_j$$

le périmètre commun entre les régions i et j . Considérant que ce facteur mesure l'imbrication mutuelle de deux régions, nous le considérons comme un paramètre de fermeture d'un point de vue Gestalt.

La propriété de continuité est plus difficile à implanter, car elle fait intervenir la combinaison de plusieurs

concepts plus haut-niveau, comme la forme et la direction générale. Nous avons choisi pour l'instant d'utiliser un critère simple pour modéliser la continuité, basé sur la forme, qui tend à privilégier le regroupement d'une petite région dans une autre :

$$CO_{ij} = \begin{cases} \varepsilon & \text{si } (N_i < NbPixMin \text{ ou } N_j < NbPixMin) \\ 1 & \text{sinon} \end{cases}$$

avec N_i le nombre de pixels de la région i . Afin de donner à CO_{ij} la même importance relative que CL_{ij} , nous posons $\varepsilon=0.25$.

A chaque itération, les régions reliées par la frontière qui présente la plus faible distance gestalt sont groupées. [9] suggère deux critères d'arrêt pour l'algorithme : toutes les distances doivent rester inférieures à un seuil maximal fixé ($MaxDG$) et le nombre total de regroupements (*gestalts*) ainsi formés doit rester supérieur à un nombre minimal fixé ($NbGestMin$).

Ainsi, l'algorithme de groupement possède un paramètre ($NbPixMin$) et deux valeurs de contrôle ($MaxDG$ et $NbGestMin$). Nous proposons que le système puisse extraire de l'ontologie du domaine les informations nécessaires pour contrôler le processus de groupement et initialiser les paramètres.

3 Ontologie des connaissances

3.1 Vers des connaissances de scène

Ici, l'ontologie doit permettre d'aider à la mise en correspondance de régions segmentées avec des concepts sémantiques. Nous proposons donc de modéliser les connaissances du domaine en connaissances de scène : l'ontologie se compose de concepts (les objets sémantiques du domaine) et de relations spatiales entre ces concepts (inclusion, adjacence et voisinage, avec différents cas pour ces deux dernières : dessus, dessous, gauche et droite). De plus, chaque objet sémantique et chaque relation sont décrits en prenant en compte leurs propriétés intrinsèques.

Enfin, comme expliqué précédemment, considérer les connaissances du domaine d'un point de vue de la scène impose de modéliser la notion de composition de scène, afin de savoir quels objets sont susceptibles d'être perçus à un niveau de détail donné.

3.2 Contrôle des *gestalts*

Puisque le processus de groupement modélise un comportement de type vision, il ne semble pas raisonnable de vouloir aboutir directement à la création d'objets sémantiques. Nous cherchons plutôt à grouper différentes régions en groupement perceptuels à forte valeur sémantique. Ainsi, il est possible que certains objets sémantiques restent décomposés en plusieurs *gestalts* distincts.

Nous proposons alors deux types de contrôle : l'un durant le processus de groupement, décrit ici et l'autre a posteriori, capable de dériver une description plus sémantique [10].

Concernant le processus de groupement, et considérant qu'il doit être effectué à chaque niveau de composition de la scène, le système peut dériver à partir des connaissances du domaine des contraintes que les *gestalts* devront satisfaire. Par exemple : nombre minimum ou maximum d'objets sémantiques, taille relative, relations spatiales, forme. Ainsi, de telles contraintes peuvent être utilisées pour contrôler le groupement (Tableau 1).

Paramètre	Valeur
$NbGestMin$	(nb d'objets nécessairement présent)*2 OU (nb d'objets susceptibles d'apparaître)*2
$NbPixMin$	(Taille minimale d'objets présents)/10

Tableau 1- Exemple d'initialisation de paramètres

La procédure pour initialiser $MaxDG$ est plus complexe : considérant que $NbGestMin$ empêche déjà le processus de faire trop d'itérations, nous suggérons d'utiliser $MaxGD$ pour privilégier le processus de groupement :

$$MaxGD = \text{mean}(S_{ij}) + \alpha \text{std}(S_{ij})$$

Avec std la déviation standard et α qui représente la granularité du groupement attendu, fixé par heuristique à chaque niveau de détail.

Concernant le contrôle a posteriori, nous proposons d'extraire de l'ontologie les contraintes géométriques que les objets sémantiques doivent vérifier, afin de pouvoir inférer une interprétation sémantique des *gestalts* [10].

Enfin, des outils génériques, capables d'extraire automatiquement de l'ontologie l'information nécessaire pour chaque contrôle, doivent être implantés. De tels outils correspondent à des opérations logiques de premier ordre, et peuvent donc être implantés en utilisant des systèmes experts qui s'interfaçent aux ontologies.

4 Exemple d'application

Notre travail est toujours en cours. Néanmoins, nous présenterons ici quelques résultats, relatifs à un exemple de domaine d'application : l'archéologie. Plus précisément, la Maison de l'Orient et de la Méditerranée (MOM) possède environ 3000 images de stèles thessaliennes, issues de la numérisation de photographies argentiques. Une stèle thessalienne est un bloc de pierre peint et sculpté, qui était utilisé dans la Grèce Antique pour marquer l'emplacement d'une sépulture.

En utilisant les descripteurs de texture de Laws [11], suivis d'un algorithme de nuées dynamiques (K-Means), nous obtenons une image segmentée composée d'environ 600 régions.

L'ontologie du domaine a été modélisée en collaboration avec les experts de la MOM et a abouti à une modélisation à base de *frames*. Ce formalisme regroupe la

connaissance d'un concept en « paquet » appelé *frame*, auquel sont associés des propriétés (*slots*), éventuellement contraintes par des facettes.

L'ontologie est composée d'environ 100 *frames* directement relatives aux connaissances du domaine, stockées dans l'éditeur d'ontologie *Protégé-2000* [12].

Les traitements génériques d'extraction de la connaissance depuis l'ontologie ont été développés en utilisant le système expert *Jess*, via un *plug-in* dans *Protégé-2000* autorisant une mise en correspondance en temps réel des *frames* de l'ontologie et des faits dans le formalisme *Jess*. Un jeu de 40 règles génériques implémente l'extraction de contraintes depuis l'ontologie et permet l'initialisation des différents paramètres de l'étape de groupement.

Des expérimentations ont été réalisées sur 20 images, désignées comme représentatives du domaine par les experts. Les images segmentées sont composées d'environ 600 régions. Après groupement au premier niveau de détail, on obtient entre 7 et 20 *gestalts*, pertinents d'un point de vue sémantique.

Le Tableau 2 présente les images segmentées et groupées en *gestalts* ($\alpha=3.5$, $NbGestMin=7$, $NbPixMin=200$) pour quatre images de stèle, au premier niveau de détail. En (a) et (b), le processus de groupement a été stoppé par le paramètre *MaxDG*, créant ainsi plus de *gestalts* qu'en (c) et (d), qui ont été contrôlés par *NbGestMin*. A ce niveau d'interprétation, trois objets étaient attendus : le couronnement, le corps, et le *geison* (fin composant entre les deux premiers). Ce dernier est absent en (c) et (d) à cause de la faible différence de ses descripteurs de texture par rapport à ceux de ses voisins. Néanmoins, le système est capable de remarquer son absence avec les contrôles a posteriori [10]. Une nouvelle analyse locale permet alors de le retrouver.

Notons également que, toujours avec l'analyse des contraintes géométriques, il est possible de regrouper une partie des *gestalts* en objets sémantiques [10]. De plus, les artefacts présents dans les corps en (a) ou (d) peuvent être écartés, de la même manière.

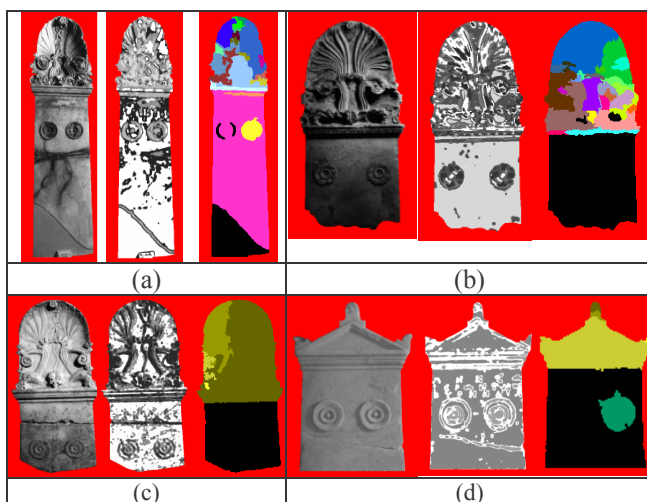


Tableau 2- Exemples de résultats

5 Conclusion et perspectives

Nous avons proposé une architecture générique intégrant les propriétés de la vision Gestalt sous le contrôle de connaissances domaine, grâce à des mécanismes flexibles d'extraction de connaissances. En outre, nous avons présenté une application de cette architecture à un domaine particulier.

Nous travaillons actuellement à une implantation plus précise des critères gestalts, particulièrement celui lié à la continuité, en utilisant un descripteur de forme.

De plus, nous projetons d'appliquer cette méthode à d'autres domaines d'interprétation, comme celui des photos aériennes.

Références

- [1] A. Smeulders, M. Worring, S. Santini et al, Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12) :1349-1380, 2000.
- [2] T. Matsuyama, V. Hang, *SIGMA : A Framework for Image Understanding Integration of Bottom-Up and Top-Down Analysis*, Plenum, New-York, 1990.
- [3] B. Draper, A. Colins, J. Brolio et al, The Schema System, *International Journal of Computer Vision*, 2(3) : 209-250, 1989.
- [4] D. Crevier, R. Lepage, Knowledge-Based Image Understanding Systems: a Survey, *Computer Vision and Image Understanding*, 67(2) : 161-185, 1997.
- [5] A. Amir, M. Lindenbaum, A Generic Grouping Algorithm and its Quantitative Analysis, *IEEE Transactions on Pattern Analysis and Machine Acquisition*, 20(2) : 168-180, 1998.
- [6] M. Wertheimer, Principles of Perceptual Organization, *Readings in Perception*, pp. 115-135, 1958.
- [7] P. Vasseur, C. Pégard, M. Mouaddib et al, Perceptual Organization Approach by Dempster-Schafer Theory, *Pattern Recognition*, 32 : 1449-1462, 1999.
- [8] A. Maßmann, S. Posch, G. Sagerer et al, "Using Markov Random Fields for Perceptual Grouping", in proc. of *International Conference on Image Processing*, Vol. 2, pp. 207-210, 1997.
- [9] K. Idrissi, G. Lavoué, J. Ricard et al, Object of Interest based Visual Navigation, Retrieval and Semantic Content Identification System, *Computer Vision and Image Understanding*, 2003, in press.
- [10] N. Zlatoff, B. Tellez, A. Baskurt, Exploitation de connaissances domaine pour l'interprétation d'images, dans *Actes de la conférence RIAO*, Avignon, 2004, in press.
- [11] K. Laws, *Textured Image Segmentation*, PhD Thesis, University of Southern California, 1980.
- [12] N. Noy, R. Ferguson, M. Musen, The Knowledge Model of Protégé-2000: combining Interoperability and Flexibility, *Knowledge Engineering and Knowledge Management: 12th International Conference EKAW 2000, Lecture Notes in Artificial Intelligence*, pp. 17-32, 2000.