Extraction et indexation d'images appliquées au domaine de la conception architecturale et technique

Jean-Claude Bignon[*] — Gilles Halin[*] — Walaiporn Nakapan[*] — Marc Wagner[**].

**UMR MAP CNRS n°694.*
CRAI - Ecole d'Architecture de Nancy
2, rue Bastien Lepage B.P. 435
54001 Nancy – Cedex
E-mail : {bignon, halin, nakapan}@crai.map.archi.fr,
***mwagner@pcpostal.com*

RÉSUMÉ. *Dans le cadre de nos travaux de recherche sur l'assistance à la conception architecturale, nous avons développé un processus de recherche par l'image pour retrouver des produits et matériaux du bâtiment. Afin de construire la base d'image pour cet outil de recherche, nous collectons régulièrement des images extraites des sites des fournisseurs de produits. Nous présentons dans cet article les techniques nécessaires à la sélection d'images pertinentes à partir des sites WEB. L'application en cours de ces techniques est un robot spécialisé à l'extraction et l'indexation d'images. Le cadre d'expérimentation de celui-ci est la base de données informatives sur les produits du bâtiment du CRIT (Centre de Ressources et d'Informations Techniques de Lorraine et d'Alsace).*

ABSTRACT. *Within the framework of our research on the architectural design assistance, we have developed an image retrieval process for the building products and materials search. In order to construct the image database for this search tool, the images appear on the product providers' web sites will be extracted and collected regularly. This article presents the necessary techniques for the pertinent image selection from the building product web site. The application in progress of these techniques is a robot specialised in image extraction and indexing. The experimentation framework is the informative building product database of CRIT (Centre de Ressources et d'Informations Techniques de Lorraine et d'Alsace).*

MOTS-CLÉS : *Extraction de connaissances, web, recherche d'informations, image, indexation semi-automatique, architecture*

KEY WORDS: *Knowledge extraction, web, information retrieval, image, semi-automatic indexing, architecture*

Signature de l'article : nom de la revue. Volume 1 – n° 1/1998, pages 1 à x

1. Introduction

Toutes les recherches en cours sur les pratiques de conception font apparaître l'importance de l'image pour les architectes. Dans une documentation sur les produits du bâtiment où en règle générale, chaque gamme, chaque produit, chaque exemple de réalisation est illustré par une image, l'utilisation de l'image comme support à la recherche d'informations semble plus que pertinente. L'efficacité de l'image à condenser l'information permet à l'utilisateur de mettre rapidement son besoin en adéquation avec l'information présentée.

Nos travaux de recherche sur la modélisation et l'assistance à la conception technique en architecture, nous ont amené à développer un processus de recherche à partir d'images permettant de retrouver des produits et matériaux du bâtiment. Cette forme de recherche devient réellement pertinente grâce à l'utilisation d'INTERNET. En effet, afin d'assurer une veille technologique constante, nous collectons régulièrement, grâce à l'utilisation d'un robot spécialisé, des images extraites des sites des fournisseurs de produits.

Nous présentons dans cet article, les critères et les techniques nécessaires à la sélection d'images pertinentes à partir de sites Web, ainsi que l'utilisation de ces images dans un processus de recherche interactif et progressif. Ces nouvelles technologies trouvent leurs applications comme mode de navigation dans la base de données informatives sur les produits du bâtiment du CRIT[1] (Centre de Ressources et d'Informations Techniques de Lorraine et d'Alsace).

2. Contexte : la recherche de produits par l'image

Le terme d'assistance à la conception renvoie à des méthodes et des attitudes différentes selon que l'on se situe à l'amont du processus (Esquisse, APS) ou à l'aval (APD, PEO,…). Afin de bien borner les limites de notre travail, il est important d'indiquer que nous travaillons actuellement plus spécifiquement sur l'assistance à la recherche d'informations relatives aux produits du bâtiment. Un tel thème concerne généralement une phrase déjà avancée du processus de conception.

Dans un tel contexte, notre hypothèse particulière est la suivante.

La recherche d'information fondée sur une approche par critères explicites (nom d'un produit, nom d'une marque, performances particulières, …) est bien adaptée à des situations où le travail de conception relève moins d'une activité de création que d'une activité de vérification (calcul de structure, thermique, CCTP,…). Elle est en revanche souvent mal adaptée à des stades antérieures lorsque le concepteur doit faire ou qu'il doit trouver des solutions répondant à des critères multiples. Dans ce type de raisonnement plus incertain, il est utile de travailler sur des modes de recherche de l'information plus flous et moins focalisés.

[1] www.crit.archi.fr

Nos travaux visent donc à mettre en place des méthodes et des outils de recherche par l'image qui utilise au mieux les potentialités de ce média et les aptitudes des architectes à raisonner à partir de figures visuelles.

Nous développons en particulier une méthode de recherche interactive et progressive d'images [HAL 90]. Elle permet à l'utilisateur d'exprimer sa demande à travers le choix d'images, en évitant la manipulation de vocabulaires précis du domaine.

Ce processus repose sur l'utilisation d'un bouclage de pertinence [RIJ 79] composé de visualisation, choix, et analyse de choix. Il a été réalisé de la manière suivante :

A. Une première fenêtre demande à l'utilisateur soit de formuler une première demande en choisissant une fonction constructive, soit d'obtenir les premières images à partir d'un tirage aléatoire.
B. Le premier ensemble d'images est présenté sous la forme d'une mosaïque d'image.
C. L'utilisateur visualise ces images (cf. Figure 1) et donne son avis sur chacune d'elle : « oui », « non », « peut-être ».
D. Après cette étape, l'utilisateur à la possibilité de continuer le processus en demandant au système de nouvelles images (E) ou de l'arrêter en demandant les produits correspondant à son choix (G).
E. Le système analyse les choix afin de construire une nouvelle requête pour sélectionner de nouvelles images.
F. Les images déjà choisies et les images les plus pertinentes du nouvel ensemble d'images sélectionné sont présentées à l'utilisateur (C).
G. L'analyse des choix permet au système de construire une nouvelle requête pour sélectionner dans la base des produits ceux qui illustrent ce choix.

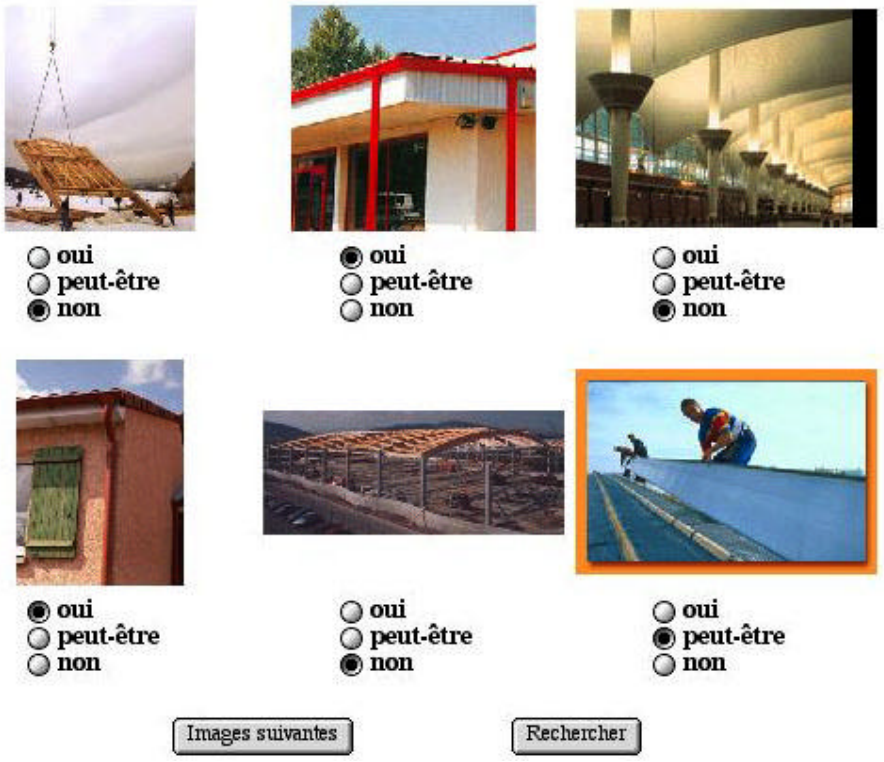

Figure 1 . *La mosaïque de la recherche par l'image*

Figure 2 . *L'image et son indexation*

Ce processus repose sur l'utilisation d'une base d'images et d'un thesaurus spécialisé dans la construction. L'information principale dans ce processus est donc l'image et son indexation (cf. Figure 2).

3. Internet, source d'images

Internet et plus particulièrement le Web représente une source d'informations où les images sont nombreuses et variées. Nombreux sont maintenant les fabricants qui proposent leur catalogue illustré de produits sur Internet. L'analyse de ces sites et l'extraction de leurs images, à l'aide de critères de sélection que nous allons présenter, permettra d'approvisionner régulièrement notre base d'image. Par cet approvisionnement régulier, nous réalisons une veille technologique par l'image, aide indispensable à tout concepteur.

Cependant, les images que l'on trouve sur le Web représentent plusieurs types. L'image ci-dessous (cf. figure 3) illustre la décomposition d'une page web d'un catalogue de radiateur. Elle est composée de cinq images :une image de titre, une icône, un dessin, une photo, et un logo. Elle figure également le contexte, qui est le paragraphe de texte illustrant le produit. Les liens hypertexte vers d'autres pages se trouvent en bas de la page.

Cet exemple montre que les images que l'on trouve sur Internet sont de plusieurs types. Toutes les images ne sont pas intéressantes pour notre approche. La question posée est donc « comment peut-on n'extraire que les images pertinentes à partir du web ? » et « comment les indexer ? ». Le paragraphe 4 présente les principes qui permettent la sélection des images pertinentes. Le paragraphe 5 présente le processus d'extraction et d'indexation d'image.

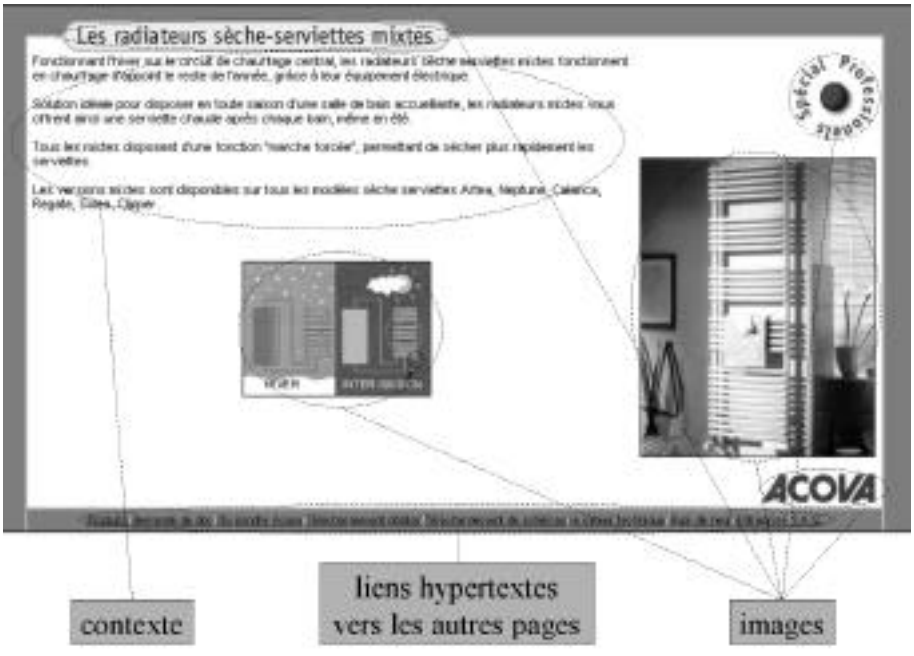

Figure 3 . *décomposition d'une page web d'un catalogue de produits*

4. Propriétés d'une image pertinente

Nous proposons trois catégories de principes permettant de déterminer qu'une image est pertinente pour la recherche d'informations techniques : les principes de reconnaissance visuelle, les principes de format, et les principes de contexte.

4.1. *Principes de reconnaissance visuelle*

L'image produit deux effets : l'effet de reconnaissance et l'effet de suggestion. L'effet engendré par l'image, qui peut nous aider à sélectionner les images visuellement, est l'effet de reconnaissance. En effet, une image est pertinente si le produit représenté dans l'image est reconnu facilement. Les critères suivants facilitent la reconnaissance de produit :

- Distinction : L'image peut représenter un ou plusieurs produits. Si elle en représente plusieurs, on doit pouvoir les distinguer les uns des autres. Ce critère s'appuie sur les théories des caractéristiques [REE 99]. Cette théorie explicite notre manière de reconnaître les formes. Une caractéristique distinctive est une caractéristique présente dans une forme, mais absente dans une autre. Elle facilite la discrimination des deux formes.
- Focalisation : Si l'image représente plusieurs produits, un seul doit dominer la présentation (par surface occupée, par contraste, par couleur, par centralisation, par répétition, …). La focalisation est une des caractéristiques de l'attention [REE 99]. Elle est nécessaire pour nous éviter d'être surchargés d'informations.
- Intégralité : L'objet doit plutôt être représenté en entier. Plus il sera tronqué, plus il faudra interpréter les parties manquantes. Cette caractéristique s'appuie sur la théorie de reconnaissance par composants de Biederman [BIE 87]. Cette théorie a relevé que nous n'avions besoin que de 35 volumes (qui ont appelés *geons*) pour décrire les objets dans l'environnement. La reconnaissance d'une forme consiste en une description de relation entre ces geons. La suppression d'information sur les relations entre ces géons réduirait la capacité des individus à reconnaître des formes.,
- Similarité de couleur : Il doit exister une similarité de couleur entre l'image et les couleurs habituellement dominantes du produit représenté.
- Environnement : Il faut maintenir dans la représentation de l'objet des éléments de son environnement d'usage. (Par exemple, un type de robinet sera mieux perçu s'il est situé à proximité d'un évier ou d'une chaudière).

4.2. *Principes de format*

JC. Vendrig propose une classification d'images sur le web dans [VEN 97]. Au total, cinq types d'images sont classifiées : photo, alphanumérique, icône, porteuse de légende, et décorative. Il s'avère que les images qui nous semblent pertinentes

pour notre recherche ne sont que les photos. Elles contiennent beaucoup d'informations. Les autres types d'images (alphanumérique, icône, porteuse de légende, et décorative) sont souvent des systèmes graphiques, pauvre en information. La taille de l'image doit être également prise en considération. Les images trop grandes (comme les images de fond), ou trop petites (comme les icônes ou les puces) ne sont pas pertinentes. La proportion de l'image aide aussi à juger sa pertinence. Une image trois fois plus longue que large est souvent une bannière ou un trait horizontal.

C'est pourquoi, des principes de format sont proposés. Ceux-ci nous permettent de sélectionner les images sans connaître leurs contenus. Ils prennent en considération le nombre de couleurs utilisées, la largeur, la hauteur, et la proportion de l'image. Une image est pertinente graphiquement si elle correspond aux critères suivants :

- l'image doit être de type photographique, c'est-à-dire riche en couleur.
- la taille d'une image sélectionnée (largeur, hauteur) doit être dans une intervalle limitée, à étudier par la statistique des images pertinentes sur le web,
- la proportion d'une image (largeur/hauteur) doit être dans un intervalle limité, proche d'une proportion d'un carré.

4.3. *Principe de contexte*

Ce principe permet de minimiser l'effort d'interprétation. Au-delà du décodage sémiologique, l'interprétation d'une image implique des processus inférentiels. Cette interprétation repose sur des informations non codées dans l'image généralement appelées contexte [REB 98]. Une image sans contexte, c'est-à-dire sans texte ou légende associée, n'est pas pertinente. Pour qu'une image vérifie le principe de contexte :

- il faut que le contexte proche de l'image ne soit pas hors sujet, il doit faire référence à des informations du domaine, ici les produits du bâtiment,
- il faut qu'il y ait un mot correspondant au terme du thesaurus dans le contexte.

5. Extraction et indexation d'images à partir du Web

Nous avons développé un processus d'extraction d'images à partir du Web (cf. Figure 4) qui comporte une sélection des images et une indexation de leur contexte. La sélection est le résultat de l'application d'une suite de critères permettant d'appliquer une partie des principes énoncés dans le paragraphe précédent. Elle est composée de deux parties principales : extraction/indexation automatiques par un robot, et validation manuelle.

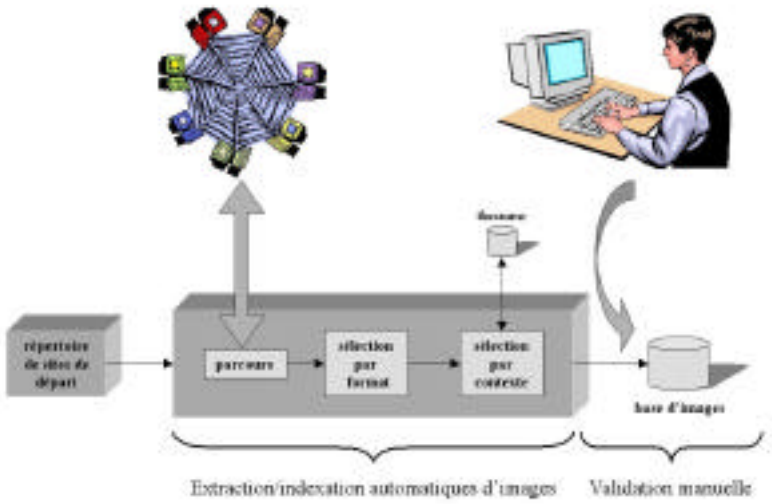

Figure 4 . *Processus d'extraction d'images*

5.1 *Sélection d'images par le robot*

L'ensemble des critères appliqué peut être représenté par une suite de questions auxquelles le robot doit répondre par l'affirmative pour sélectionner une image. Les questions A, B, C, E et F proviennent des principes de contexte. La question D provient des principes de format. Notez que les principes de reconnaissance visuelle ne sont pas appliqués ici. Un contrôle de pertinence visuelle par un administrateur est nécessaire. La sélection d'images pertinentes est effectuée automatiquement par le robot et contrôlée manuellement par l'administrateur. C'est pourquoi notre processus est semi-automatique.

Chaque question est présentée d'une manière suivante :

A. *« La page où se situe l'image est-elle à une distance proche de la racine du site parcouru ? »*

La distance des pages web par rapport à leurs parents doit être vérifiée afin d'éliminer les pages qui ne concernent pas les produits du bâtiment. Comme nous partons des pages d'accueil ou des pages d'index, les pages trop éloignées de celles-ci ont une forte chance de représenter autre chose et elles seront rejetées. Il y a deux notions de distance. Lorsqu'une distance est égale à 0, qui est la distance par défaut de parcours, on parcourt tous les répertoires du même domaine ayant les mêmes

URL parentaux. Lorsque la distance est égale à 1, qui est la limite de parcours, seuls les domaines référencés directement par les URL parentaux sont acceptés et pas plus loin.

B. *« La page où se situe l'image est-elle en français ? »*
Le choix du langage utilisé dans la page peut un critère discriminant des pages écrites en langues étrangères. Puisque le thesaurus utilisé est en français, ce n'est pas possible d'indexer les images provenant de ces pages-là.

C. *« L'image est-elle dans une page intéressante (présentation de catalogue) ? »*
Les images intéressantes se trouvent sur les pages présentant les produits du bâtiment, dans lesquelles les images et leurs contextes sont très pertinents. Au contraire, les images présentées sur la page « présentation de la société » ou « contacte » ne sont pas pertinentes parce qu'on ne trouve pas la présentation de produit. Ce critère nous aide à éliminer les images parasites comme la carte de France, le plan d'accès, la photo de l'usine, etc.

D. *« Le format de l'image est-il bon ? »*
Ce critère examine l'aspect physique de l'image en considérant sa dimension (largeur, hauteur), sa proportion (largeur/hauteur), et sa taille en octet. Nous pouvons éliminer les images décoratives comme les puces (15*15 pixels^2) ou les bannières (3 fois plus longue que large). La taille en octet nous dit si l'image est assez grande et si ça vaut la peine de la vérifier. Voici les critères validés qui sont résultats des statistiques étudiées par [NAK 98] :

- 60 pixels ≤ largeur ≤ 610 pixels,
- 60 pixels ≤ hauteur ≤ 660 pixels,
- 0,58 ≤ proportion ≤ 2,1,
- 3,2 kilooctet ≤ taille en octet.

E. *« L'image a-t-elle un ou plusieurs contextes ? »*
L'image est rejetée à cette étape si on ne trouve pas un contexte proche. Cependant, le contexte d'une image peut être le contexte d'une autre. La distance entre le contexte et l'image nous aide à décider quel contexte appartient à quelle image.

F. *« Le contenu du contexte est-il intéressant ? »*
Même si on trouve un contexte proche de l'image, cela ne sert à rien s'il n'est pas intéressant. Pour notre approche, un contexte est intéressant s'il contient un terme équivalent du thesaurus de produit du bâtiment.

Le résultat de l'extraction est une liste d'images associées à leurs contextes.

5.2 *Indexation*

L'indexation procède par l'analyse des contextes extraits précédemment, afin de déterminer les termes du thesaurus qui feront partie de l'indexation de l'image. Le processus d'indexation s'appuie sur la technologie des n-grammes [HAL 99], il suit les étapes suivantes :

- à chaque terme du thesaurus est associée sa représentation en n-grammes,

- les contextes de chaque image sont analysés afin d'en extraire des groupes nominaux. À chacun de ces groupes nominaux est alors associée sa représentation en n-grammes,
- une fonction de mise en correspondance évalue la distance, à l'aide des représentations en n-grammes, entre chaque groupe nominal contenu dans les contextes et les termes du thesaurus.
- un tri est alors effectué pour sélectionner les termes du thesaurus les plus pertinents.

On obtient ainsi pour chacune des images extraites une indexation contenant un vecteur pondéré de termes du thesaurus. Ces images et leur indexation vont être le support au processus de recherche interactive et progressive d'images.

6. Application : Wimex-bot

Un des premiers travaux menées en matière d'indexation fût le *Marie project* de US Naval Postgraduate school [ROW 97]. L'application utilise le code source HTML afin d'indexer l'image en utilisant la légende. Un autre travail qui s'y rattache est *Image Excavator* de School of Computing Science, Simon Fraser University, Canada [ZAI 98]. Le système utilise les informations textuelles, comme les balises HTML, afin de dériver les mots clés en anglais. En traversant les structures en ligne comme Yahoo !, il peut créer les hiérarchies des mots clés. Ces derniers sont mis en correspondance avec les termes du répertoire dans lequel l'image est trouvée.

Notre robot, Wimex-bot, est un robot spécialisé qui utilise les techniques décrites ci-dessus (paragraphe 5.1 et 5.2). Programmé en Java, il extrait et indexe les images de produits du bâtiment à partir de code source HTML. Il peut approvisionner et mettre à jour la base d'image pour la recherche de produit du bâtiment. À partir d'une liste des pages, qui sont souvent les pages d'accueil, ou la page d'index, le robot traverse le web. Ensuite, il analyse la page afin de trouver les pages suivantes. Enfin, il extrait les images et les indexe en utilisant les termes du thesaurus.

L'interface graphique permet à l'utilisateur de visualiser l'avancement du traitement par le « spider ». Il peut également parcourir l'arborescence des sites web analysés. Ensuite, il peut visualiser les informations correspondantes à son choix de liens hypertextes (page, image) dans la phase de validation des images.

Lorsque l'utilisateur choisit un lien d'une page, la liste d'images extraites par le robot apparaît dans l'arborescence (cf. figure 5). Par conséquent, il peut visualiser et valider l'image qu'il trouve pertinente. Lorsqu'il choisit une image, il visualise le résultat de l'indexation. Cela lui permet de valider ou rejeter l'image. S'il trouve que l'indexation est incomplète, il peut la compléter en choisissant un ou plusieurs termes à partir du thesaurus.

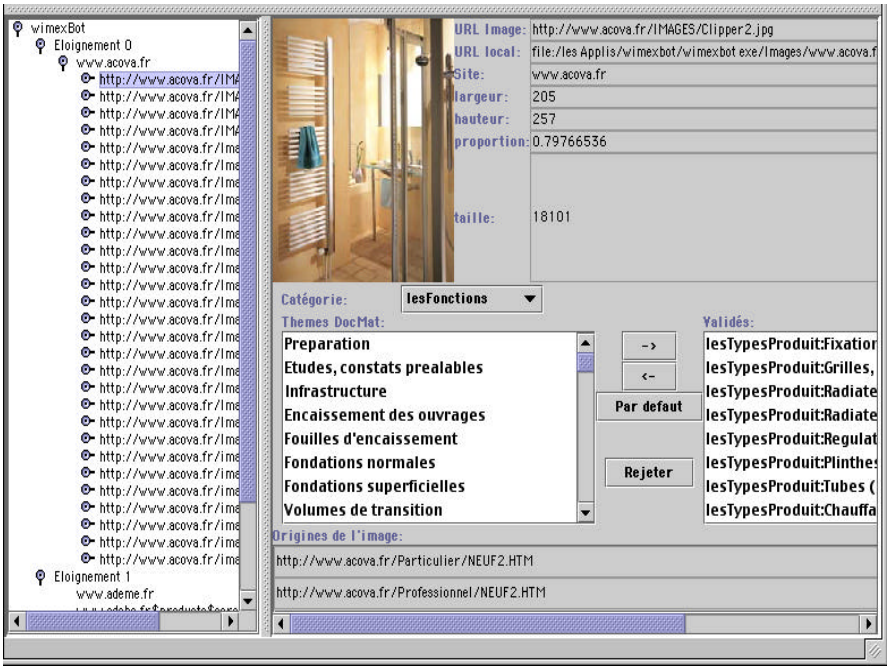

Figure 5 . *le résultat d'extraction*

Afin de tester notre robot, nous avons mené une expérimentation sur les 6 sites de fabricants suivants :

- http://www.atlantic.tm.fr
- http://www.godin.fr
- http://www.fluidest.com
- http://www.paul-mathis.fr
- http://www.alutherm.fr
- http://www.nailweb.com

Memoweb :	*Wimex-bot :*
Nombre total d'images : 318 images extraites 48 images pertinentes	Nombre total d'images : 35 images extraites (sur 318) 24 images pertinentes (sur 48)
⇩ 11 bruits (32%) 24 silences (50%)	

Tableau 1. *L'exemple de l'analyse du résultat du site « atlantic »*

Afin de pouvoir mener une comparaison nous avons utilisé le logiciel « Memoweb »[2] pour aspirer les images et les pages des sites analysés. Par exemple pour le site « atlantic », Memoweb a extrait 318 images, soit la totalité des éléments de type image. 48 images ont été considérées par un expert comme pertinentes après application des principes de format et de reconnaissance visuelle. Pour le même site, Wimex-bot a extrait 35 images après application des principes de format et de contexte dont 24 ont été reconnues pertinentes selon les principes de reconnaissance visuelle.

Au total le résultat d'extraction des 6 sites représente pour Memoweb 1509 images dont 315 sont pertinentes et pour Wimex-bot 149 images dont 53 sont pertinentes. Soit 83% de silence et 64% de bruits.

Après analyse de l'indexation fournie par Wimex-bot, il apparaît que 65 % des images indexées sont incorrectes (le terme ne correspond pas à l'image) ou incomplètes (toutes les catégories du thesaurus ne sont pas utilisées) ou incohérentes (plusieurs termes de produits différents désignent le même produit).

7. Conclusion

Il résulte de cette expérience plusieurs éléments de conclusion :

- le nombre des images extraites par rapport aux images disponibles sue les sites est beaucoup plus faible. Dans une situation pléthorique d'images sur le web, ceci est plutôt satisfaisant.
- le taux de pertinence (la précision) augmente (de 315 sur 1509, soit 20,87 % à 53 sur 149, soit 35,57 %) mais il reste insatisfaisant.
- La qualité de l'indexation permet déjà un premier niveau d'automatisation mais reste faible (65 % de « mauvaise » indexation).

Pour améliorer notre méthode et notre robot, nous envisageons donc plusieurs pistes :

- une adaptation du thesaurus qui reste mal adapté à l'indexation d'éléments visuels.
- L'implantation des techniques de traitements d'images comme l'identification de forme par vecteur de caractéristique, par relation entre les éléments, ou par transformation de forme [DEL 99].

8. Références

[BIE 87] Biederman I., « Recognition-by-Components: A Theory of Human Image Understanding », *Psychological Review*, n° 94, p. 115-147, 1987.

[2] http://www.goto.fr/fr/news/mwus3002.htm

[DEL 99] Del Bimbo A., *Visual Information Retrieval*. Morgan Kaufmann Publishers, San Francisco, California, 1999.

[HAL 90] Halin, G., Créhange M., Kerekes P., « Machine learning and vectoriel matching for an image retrieval model: EXPRIM and the system RIVAGE ». *Proceedings of the ACM 13th International Conference on Research and Development in Information Retrieval SIGIR'90*, Brussels, 5-7 September 1990, p. 99-114.

[HAL 99] Hallab M., Lelu A., « Proxilex : un outil d'approximation orthographe à partir des fréquences des n-grammes ». *Hypertextes hypermédias et internet, 5e conférence internationale H2PTM'99*, Paris, 23-24 septembre 1999, p. 201-209.

[NAK 98] Nakapan W., « Navigation thématique par l'image. Extraction et indexation semi-automatique d'image à partir du Web ». Mémoire de DEA. Université de Nancy 1, 1998.

[REB 98] Reboul, H., Moescler J., *La Pragmatique aujourd'hui. Une nouvelle science de la communication*, Seuil, Paris, 1998.

[REE 99] Reed S., *Cognition. Théories et applications*, De Boek Université s.a., Bruxelles, 1999.

[RIJ 79] van Rijsbergen C.J., *Information Retrieval*. 2nd edition Butterworths, London, 1979.

[ROW 97] Rowe N. C., Frew B., « Finding photograph captions multimodally on the World Wide Web ». *Technical report from the AAAI Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*, p. 45-51, 1997.

[VEN 97] Vendrig J., « Filter Image Browsing, A study to image retrieval in large pictorial databases », Thesis, Faculty WINS, Universiteit van Amsterdam, 1997.

[ZAI 98] Zaïane, O.R., Han, J., Li, Z.-N., Hou, J., « Mining Multimedia Data », *Proceeding of CASCON'98: Meeting of Minds*, Toronto, Canada, p. 83-96, November, 1998.

9. Annexe

Image Extraction and Indexing Applied to the Domain of Architectural and Technical Design.