Indexation et Recherche d'Images

Roger Mohr[(1)], Patrick Gros[(2)], Bart Lamiroy, Sylvaine Picard, Cordelia Schmid[(3)]

(1) INPG (2) CNRS (3) INRIA
projet MOVI, GRAVIR–IMAG
INRIA, 655 av. de l'Europe
38330 Montbonnot
email: nom.prenom@imag.fr

RÉSUMÉ

Cet article présente les problèmes et les amorces de solutions posés par la recherche d'images dans des bases d'images dès lors que l'on souhaite une indexation automatique, comme c'est le cas pour des documents écrits. Les outils utilisés pour caractériser des images et permettre une recherche de ce type sont actuellement frustes, et souvent reposent sur des caractéristiques globales utilisant largement l'information de luminance. Nous plaidons pour l'usage de caractéristiques locales bien qu'ils aient l'inconvénient de se heurter à des problèmes de segmentation. Nous montrons, par un exemple détaillé, que des éléments de solution existent et qu'ils peuvent indiquer la voie pour des recherches futures. Cet article illustre aussi l'intérêt que présentent les déjà anciennes recherches en reconnaissance des formes pour ce genre de problèmes.

ABSTRACT

This paper presents the problems that arise when retrieving images from a large image base by an automatic indexing process, as done with written documents. The tools used to characterize the images and retrieve image data are still sketchy, and usually use global features and illumination information. We plead for the use of local characteristics, even though they have the drawback of requiring a segmentation stage. We show by a detailed example that some elements of solutions already exist, and that they show some directions for future research. This paper also illustrates the interest of the classical results in pattern recognition in this kind of problems.

1 Le problème

1.1 Le besoin

La recherche d'information par le contenu est effective dans des documents écrits. En revanche, l'équivalent pour l'image en est encore à un stade rudimentaire. La raison majeure en est évidente : les mots d'un texte convoient une information symbolique proche d'un utilisateur qui peut demander des « textes parlant de poésie dans un contexte de nature » ; on imagine difficilement une requête équivalente pour des images.

Pourtant le besoin existe de faciliter l'accès aux images. Un premier besoin vient de la multiplicité de ces dernières : un simple clic permet de les créer, et, de plus en plus, le format numérique permet leur manipulation et leur diffusion aisées. Pour se rendre compte du volume que cela représente, il suffit de penser aux Giga-octets d'images que les satellites déversent sur les stations réceptrices, ou aux millions d'images que les agences de presse archivent. Ensuite, l'image reste un moyen de communication privilégié : le contenu d'une figure, d'un dessin peut difficilement se transposer dans le langage naturel. On souhaite donc pouvoir avoir accès naturellement à ce type de données pour les besoins de documentation.

Les systèmes actuellement disponibles utilisent des indexations manuelles. Ils sont parfaitement bien adaptés dès lors que le nombre d'images reste raisonnable (quelques dizaines de milliers) et que la nature de leur contenu informatif est parfaitement déterminé. Or l'image, le plus souvent, par son contenu très riche, dissimulera des informations non codées à l'avance et donc limitera la portée d'une indexation manuelle. De plus, dans certains cas, il est certain que l'information ne pourra se faire à travers des mots clés alors que l'image fournira elle même une *signature* identifiable ; par exemple, cela peut être le cas pour une texture. Enfin il est des applications où le volume d'information à gérer dépasse le raisonnable pour une indexation manuelle. Cela lance donc le défi d'une automatisation de l'indexation d'images par le contenu.

Ce problème est, à première vue, effrayant. Remarquons cependant une note optimiste : à la différence de la consultation de documents écrits, l'analyse d'une image par un utilisateur peut se faire en un temps très court, de l'ordre de la seconde. On pourra donc se permettre d'avoir des réponses plus nombreuses et plus approximatives. Il faudra en revanche imaginer un système d'interaction intelligent pour permettre à l'utilisateur de cheminer dans ce large ensemble de réponses.

De fait, quelques systèmes répondent déjà à ce besoin émergent. Ce sont des systèmes avec des capacités limitées et ils seront discutés à la section 5. Les sociétés qui les commercialisent clament fortement la réalité de ces besoins, et ce malgré l'état embryonnaire des fonctionnalités qui sont offertes.

1.2 La problématique scientifique

Il y a vingt ans et plus, la reconnaissance des formes était un sujet particulièrement actif. Le principe était pour l'essentiel de classer des formes à partir de mesures qui commençaient à être prises avec des caméras. Le schéma était pour l'essentiel le suivant :

- obtenir des mesures qui soient invariantes par rapport à la variabilité de l'observation (par exemple des moments centrés)
- déterminer la similarité de cet ensemble de mesure avec des échantillons obtenus lors d'une phase d'apprentissage.

Le problème posé possède une certaine similarité : il faut retrouver les images semblables à celles que l'utilisateur recherche. La difficulté est, d'une part, l'expression de cette fonction de similarité, et, d'autre part, la capacité à exprimer « l'objet » de la similarité, un concept qui n'est typiquement présent que dans l'esprit de la personne tentant d'émettre une requête. Ce problème est souvent contourné en demandant à l'utilisateur d'exprimer ce qu'il recherche en utilisant des images exemples. Supposons, pour simplifier, que la requête soit réduite à « trouver les images qui contiennent des choses similaires à cette partie là de cette image ». Le problème se concentre sur l'objet à comparer, mais reste entier sur la fonction de comparaison : parle-t-on de texture ? d'un objet précisément présent dans l'image ? d'un concept plus abstrait comme l'existence d'une statue ? Pour permettre de déterminer cette fonction, il faudra donc imaginer une interaction qui devra limiter la nature de la fonction. On pourra ainsi proposer des réponses et analyser l'attitude de l'utilisateur (non pas cela mais plutôt cela), ce qui s'appelle dans le monde de la recherche d'information le *relevance feedback*.

Ces fonctions à déterminer seront évaluées à partir de mesures dans les images. Comme la reconnaissance des formes nous l'a appris, ces mesures doivent être invariantes. Invariantes par quoi ? Cela dépendra de l'application bien sûr. Mais on aimerait que la plupart du temps qu'un cadrage peu différent conduise à des résultats similaires. Parfois, on exigera une invariance simplement au groupe des déplacements dans les images. On souhaitera aussi résister à des variations d'éclairage. Il faut bien comprendre que dès que l'on observe des images qui sont des projections du monde tridimensionnel, il n'existe plus d'invariant ni géométrique ni de luminance (voir par exemple [3]). Mais en revanche, si l'on considère un développement au premier ordre des variations géométriques, le groupe des similitudes capture bien les variations géométriques dans une image, et on parlera dans ce cas de quasi-invariants [2]. De même, les transformations affines de la luminance capturent bien les variations d'illumination tant que les ombres portées ne perturbent pas la saisie [21].

Enfin, il faudra inévitablement absorber dans l'ensemble du processus des problèmes de bruit, mais surtout d'occultation partielle et de changement de plan. Et seules des méthodes robustes (au sens statistique de ce terme [18]) pourront atteindre cet objectif.

On peut donc résumer ce propos en disant que les problèmes à résoudre sont les suivants :

- déterminer comment l'utilisateur précisera ce qu'il désire extraire comme images ; ce problème difficile est trop dépendant du contexte et nous supposerons qu'il s'agira essentiellement de retrouver une image proche d'une image donnée durant cet article ;
- déterminer quelles mesures invariantes il faut considérer pour cela dans l'image ;
- déterminer une fonction de ressemblance qui soit statistiquement robuste.

Dans la suite nous présenterons quelques travaux face à ces critères (section 2 et 3). Nous dégagerons en détails une solution (section 4) et nous discuterons l'ensemble de ces points à la lumière de quelques expériences et produits commerciaux.

2 Indexer par des mesures globales

De nombreux systèmes d'indexation utilisent des informations globales. Ils considèrent donc les images soit globalement, soit en fonction d'un découpage a priori (en 4, 9 ou 16 parties), en appliquant alors des techniques globales à chacune de ces parties, soit utilisent des caractéristiques globales, telle une silhouette d'un objet extraite par une méthode qui suppose que l'objet est clair sur fond sombre par exemple.

Trois types d'informations sont principalement utilisés. Tout d'abord la couleur, ensuite les niveaux de gris, qui sont soit utilisés comme la couleur, soit sous forme de texture, et enfin les silhouettes sus-citées.

Ces méthodes permettent des résultats attrayants au premier regard, car elles sont basées sur des critères globaux tels la couleur qui ont une grande importance visuelle pour le premier coup d'œil à une image. Elles sont donc adaptées aux tâches de présélection d'images ayant des caractéristiques dominantes : couchers de soleil sur la mer, textures et couleur de forêt... Par contre, ces méthodes sont très sensibles à tous les changements qui font varier la composition de l'image. Ainsi, une translation de l'image, qui fait perdre une partie du signal et fait apparaître une région nouvelle, un changement du fond sur lequel apparaît l'objet ou une occultation partielle ont des effets catastrophiques sur les performances de ces méthodes, qui se révèlent donc inaptes à la reconnaissance d'objets.

2.1 Utilisation de la couleur

Une première manière d'utiliser simplement une information de couleur est de rechercher la couleur dominante dans l'image ou dans une partie de l'image. Cette approche peut paraître rustique, mais elle peut fournir un moyen efficace pour formuler des requêtes pour qui a une idée visuelle vague de l'image qu'il recherche : « il y avait une fleur rouge dans le coin en bas à gauche sur un ciel bleu... »

Une manière plus fine d'utiliser la couleur consiste à calculer des histogrammes. Pour cela, l'information est décom-

posée en trois composantes, RGB ou selon un autre système faisant intervenir la luminance globale, puis un histogramme tridimensionnel est calculé. Pour comparer deux images, on compare alors leur histogramme. Diverses méthodes sont envisageables pour cette comparaison [25, 24, 14, 30].

L'utilisation d'histogrammes mène à trois inconvénients majeurs : ce sont des données globales qui souffrent donc des défaut déjà signalés ; ils ne sont pas invariants au changement d'illumination (certaines représentations permettent de pallier partiellement à cela) ; enfin, la manipulation d'histogrammes ayant beaucoup de dimensions mène à des calculs de distance complexes.

Pour palier à ce dernier inconvénient, diverses méthodes ont été testées. Réduire le nombre de dimensions peut mener à des résultats instables [30]. Il a été proposé d'utiliser les histogrammes des 3 couleurs séparément [8], ou de ne calculer les distances qu'entre les moyennes de chaque couleur [9]. Une autre voie consiste à chercher de meilleurs index[6], ou les couleurs permettant la meilleure reconnaissance pour un ensemble d'images données [14].

Afin d'éviter certains problèmes liés aux histogrammes, il a été proposé [24] d'utiliser des moments. Il ne s'agit pas alors de calculer de super-couleurs, mais au contraire de calculer des quantités caractéristiques de la forme de l'histogramme. La difficulté est de déterminer l'ordre des moments à calculer. Les moments d'ordre peu élevé résument bien l'information la plus utile, mais ils ne sont pas invariants aux changements d'illumination.

D'autre part, pour assurer une véritable invariance aux variations d'illumination, il est nécessaire d'avoir un modèle de ces variations. Un modèle courant est le modèle diagonal : lors d'une variation d'illumination tous les niveaux de rouge de l'image sont multipliés par une constante, tous les niveaux de vert par une autre constante, et tous les niveaux de bleu par une troisième constante. Ce modèle n'est évidemment valide que tant qu'il n'y a pas de saturation des capteurs, il correspond à une première approximation d'une réalité plus complexe.

L'utilisation d'angles de couleurs permet d'obtenir une invariance selon ce modèle [4]. Chaque image est représentée par 3 vecteurs, un vecteur pour chaque couleur. Chaque vecteur comporte une dimension pour chaque pixel. Une variation d'illumination se traduit alors par un simple changement d'échelle du vecteur, sans changement d'orientation. Chaque image est alors caractérisée par les angles formés entre ces trois vecteurs.

2.2 Utilisation des niveaux de gris

Les niveaux de gris peuvent être utilisés de plusieurs façons différentes. Une première méthode, qui utilise de nombreux travaux passés, consiste à calculer des indices de texture pour l'image entière ou une partie de l'image. Cet indice peut être un simple indice de granularité. Comme il est souvent insuffisant à lui seul, on le trouve lié à d'autres dans des systèmes multi-critères (voir 5).

Une deuxième approche est appelée image propre ou *eigenimage* en anglais. Dans ce cas, on cherche à caractériser plusieurs sous-ensembles d'images, puis on classe alors les nouvelles images dans un des sous-ensembles précédents. Pour caractériser un sous-ensemble d'images, on calcule l'image moyenne (en calculant la moyenne des niveaux de gris pour chaque pixel), puis, par analyse en composantes principales, ou par calcul des vecteurs propres de la matrice de corrélation, on détermine la variation du sous-ensemble d'images autour de l'image moyenne.

La reconnaissance, qui est, plus exactement, une classification, est alors réalisée par projection sur chacune des images moyennes. La distance utilisée tient alors compte des variations constatées autour de chacune de ces images propres.

Un tel système a été utilisé, par exemple, pour déterminer si la route devant un véhicule allait à droite, à gauche ou tout droit, et montrait des résultats proches de ceux obtenus avec un réseau de neurones [26].

Cette approche souffre du défaut commun à beaucoup de méthodes globales. Elle est peu robuste au décalage de l'image qui fait perdre une partie de l'information.

2.3 Utilisation des silhouettes

Le dernier système [15, 16] utilisant globalement l'image se base sur la silhouette des objets. Cette méthode pourrait être jugée locale si cette extraction n'était pas réalisée par seuillage dans l'image, pour des images présentant uniquement des objets clairs sur des fonds noirs au bruit près, ce que nous ne considérons pas être une véritable segmentation.

Pour chaque objet de la base, de nombreuses images de l'objet sont prises sous différents angles. Les silhouettes extraites sont alors caractérisées par une méthode basée sur une analyse en composantes principales.

La principale limitation de cette méthode vient des conditions expérimentales imposées : l'objet observé doit être clair, sur un fond sombre, et ne pas être partiellement occulté. Ces conditions peuvent être réunies pour un système industriel, opérant en environnement contrôlé, mais ne peut plus être utilisé dès qu'il y a changement de fond, plusieurs objets présents ou occultation partielle.

Un autre exemple travail sur les silhouettes est rencontré avec [23]. L'ambition de l'auteur est de travailler avec des déformations non rigides des objets. La base des images est représentée par un certain nombre de formes prototypes. Les images de la base sont référencées par rapport à ces formes. Ce travail est intéressant, mais n'aborde que de façon très détournée le principe de l'indexation : à ce jour, il n'est pas envisageable d'avoir une indexation par le contenu avec cette approche.

3 Indexer par des mesures locales

Pour palier aux inconvénients des méthodes précédentes, de nombreuses méthodes utilisent plutôt des informations locales. Ces informations sont la plupart du temps obtenues par segmentation de l'image, que cela soit en points, contours,

segments de droites, coniques ou courbes, puis chaque primitive extraite est caractérisée par un vecteur qui sert d'index.

Cette segmentation permet d'extraire de l'image des éléments qui sont assez robustes aux changements d'illumination, ou de points de vue, ce qui est un critère important. Par contre, la précision des extracteurs n'est pas toujours satisfaisante, et il faut toujours compter avec des primitives qui ne sont pas retrouvées dans toutes les images. Il s'agit là du principal point faible de ces méthodes.

Le deuxième point délicat est le choix des caractéristiques calculées à partir des primitives extraites. Pour assurer la robustesse du système d'indexation, ces caractéristiques devraient, dans l'idéal, être invariantes à toutes les variations que peuvent subir les images : changement de points de vue, changement d'illumination. Même dans les cas les plus simples où l'on connaît de véritables invariants, comme par exemple pour les images d'une scène plane sous illumination constante, ceux-ci sont relativement complexes, et leur robustesse au bruit n'est pas assurée.

D'où l'idée, lancée indépendamment par Binford [2] et Ben Arie [1], de ne pas utiliser de vrais invariants, mais des quantités localement invariantes que Binford dénomme quasi-invariants, et dont Ben Arie prouve la stabilité tant qu'on reste loin de situations critiques. L'angle est ainsi un quasi-invariant projectif.

Cette démarche de recherche de quasi-invariants peut être étendue à toutes les variations, et permet une grande robustesse dans la pratique.

3.1 Indexation d'images structurées

À partir de ces idées, nous avons développé un système d'indexation et de reconnaissance pour les images structurées [12]. On commence par extraire les contours de toutes les images, puis ces contours sont approchés par des segments, segments qui sont reliés entre eux pour former une structure de graphe.

On utilise alors comme quasi-invariants les angles et rapports de distance pour certaines configurations de segments. Les vecteurs de caractéristiques sont stockés dans une structure arborescentes qui permet un accès rapide.

Lorsqu'on dispose d'une nouvelle image, on la traite de la même façon. Ses vecteurs de caractéristiques sont comparés à ceux de la base. Un système de vote est alors utilisé pour choisir l'image de la base qui est la plus susceptible de représenter le même objet que l'image requête.

Un simple vote comptant le nombre de vecteurs de caractéristiques presque égaux ne suffit pas, car les caractéristiques utilisées ne sont pas assez discriminantes. Chaque correspondance entre vecteurs de caractéristiques fournit une approximation du mouvement apparent entre les images. Partant de cette constatation, nous ne comptons comme vote que les correspondances qui fournissent des approximations cohérentes entre elles du mouvement apparent.

Cette technique permet un taux de reconnaissance supérieur à 92 %. La figure 1 montre un exemple des résultats que l'on obtenir. À partir de l'image en niveau de gris montrée à gauche, on obtient la structure de segments montrée au milieu. Il faut noter que cette structure est très bruitée. L'objet reconnu est montré à droite. Une fois un premier objet reconnu, il est possible de supprimer les segments correspondant à ce premier objet, puis de relancer une requête avec le reste de la structure pour reconnaître les autres objets présents.

4 Une approche par invariants de niveaux de gris

Dernièrement, nous avons proposé une nouvelle méthode d'appariement et d'indexation d'images, basée sur l'utilisation d'invariants locaux calculés à partir des niveaux de gris, mais qui suit un schéma très proche de celle qui vient d'être présentée au paragraphe 3.1. Cette section donne un aperçu de la méthode. Pour des références plus complètes, on pourra se reporter à [21, 22].

4.1 Principe de la méthode

Les trois points principaus de la méthode sont l'extraction de points d'intérêt et leur caractérisation, leur mise en correspondance, et leur utilisation pour indexer. Dans cette section nous décrivons ces trois étapes.

4.2 Extraction de points d'intérêt

L'extraction des points d'intérêt est la première étape du processus de mise en correspondance.

Pour ce faire, il existe de nombreux algorithmes dans la littérature. Dans un contexte de mise en correspondance, il est important que le détecteur soit répétable, c'est-à-dire invariant aux transformations images. Dans [21], nous avons comparé différents détecteurs en présence d'une rotation image, d'un changement d'échelle, d'un changement de luminosité et du bruit image. Les meilleurs résultats ont été obtenus avec le détecteur de Harris [7], qui cherche les points directement à partir des niveaux de gris, sans passer par une extraction des contours préalable, ce qui renforce sa robustesse.

Ce détecteur considère un point d'intérêt comme un point lequel les courbures principales de la fonction d'autocorrélation est maximale, ce qui revient à considérer les valeurs propres de la matrice $\begin{pmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{pmatrix}$ où I désigne la fonction de luminance de l'image. Une implémentation stabilisée a été utilisée dans ce travail en utilisant des dérivées Gaussiennes. Une implémentation récursive des ces filtres assure une détection rapide. Même si tous les points ne sont pas détectés de manière répétable, la proportion de ceux qui le sont suffit à assurer le succès des étapes suivantes.

4.3 Caractérisation locale du signal

De nombreuses caractérisations locales du signal sont possibles : dérivées, filtres de Gabor, ondelettes, moments. Dans ce travail nous avons utilisé des dérivées.

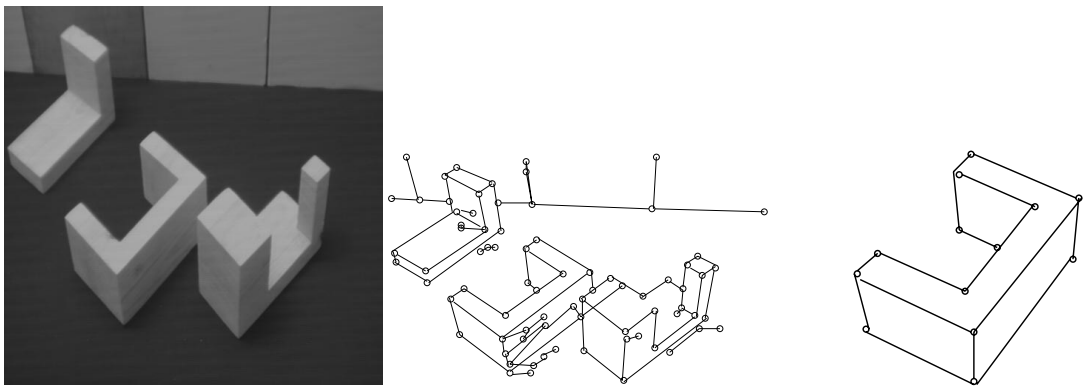

FIG. 1 — L'image à reconnaître à gauche, les segments extraits au milieu, un objet reconnu à droite.

Une fonction peut être approchée localement par ses dérivées en utilisant un développement de Taylor. De ce fait, il est possible de décrire une image en un point en stockant dans un vecteur l'ensemble des dérivées en ce point. Un tel vecteur a été utilisé par Koenderink [11] qui l'a nommé « jet local ».

Koenderink calcule, en outre, les dérivées d'une manière stable en utilisant un filtre passe-bas, c'est à dire en convoluant le signal avec les dérivées d'une Gaussienne. Le σ de la fonction Gaussienne détermine la quantité de lissage effectué. Ce σ peut également être interprété comme un facteur d'échelle. Par la suite nous appellerons σ la taille de la Gaussienne.

Une fois ces dérivées calculées, il s'agit de trouver une caractérisation qui soit invariante à différentes transformations images : rotation image, changement d'échelle et changement de luminosité.

4.3.1 Invariance à une rotation image

À partir de ces dérivées, plusieurs auteurs [11, 19, 5, 17, 27] proposent de calculer des invariants différentiels pour le groupe des déplacements $SO(2)$, en reprenant des résultats déjà formulés par Hilbert.

L'ensemble d'invariants que nous utilisons est regroupé dans un vecteur noté $\mathcal{V}$, qui est constitué d'un ensemble complet d'invariants différentiels jusqu'au troisième ordre.

La formulation de ce vecteur est donnée en notation d'Einstein (voir l'équation 1).

$$\mathcal{V} = \begin{pmatrix} L \\ L_i L_i \\ L_i L_{ij} L_j \\ L_{ii} \\ L_{ij} L_{ji} \\ \varepsilon_{ij}(L_{jkl} L_i L_k L_l - L_{jkk} L_i L_l L_l) \\ L_{iij} L_j L_k L_k - L_{ijk} L_i L_j L_k \\ -\varepsilon_{ij} L_{jkl} L_i L_k L_l \\ L_{ijk} L_i L_j L_k \end{pmatrix} \quad (1)$$

où ε_{ij} représente le tenseur canonique anti-symétrique : $\varepsilon_{12} = -\varepsilon_{21} = 1$ et $\varepsilon_{11} = \varepsilon_{22} = 0$.

En notation d'Einstein, un indice i signifie la sommation des dérivations par rapport à l'ensemble des variables :

$$L_i = \sum_i L_i = \frac{\partial L}{\partial x} + \frac{\partial L}{\partial y}$$

On peut constater que la deuxième composante de ce vecteur est la magnitude du gradient et la quatrième le Laplacien.

Les dérivées sont calculées comme indiqué précédemment, en prenant en compte le rapport d'aspect des pixels pour assurer l'invariance à la rotation. Il est possible, de plus, de calculer les L_i et donc les invariants pour plusieurs tailles de σ de la Gaussienne.

4.3.2 Robustesse à l'échelle

De même que précédemment, on peut calculer des caractérisations locales invariantes au changement d'échelle dans l'image, en prenant par exemple des rapports de dérivées. Cela dit, de tels invariants doivent être calculés sur un support, qui lui est indépendant de l'échelle. Ce support de taille fixe fait qu'on ne prend pas en compte la même information dans le deux images, et, qu'en pratique, les invariants ne sont pas stables.

Pour obtenir une caractérisation robuste à un changement d'échelle, il est donc nécessaire d'utiliser une approche multi-échelle. On calcule simplement les invariants sur des supports de différentes tailles avec le σ de la Gaussienne qui varie en conséquence. Cette approche suit celle déjà proposée par plusieurs auteurs [28, 10, 13]. Pour le choix des valeurs, de nombreux auteurs ont proposé une discrétisation par demi-octave [29]. Avec un tel pas de discrétisation, la caractérisation obtenue s'est révélée imprécise et instable. Puisque notre caractérisation s'est avérée robuste à un changement d'échelle jusqu'à 20%, nous avons choisi un pas de discrétisation qui garantit qu'entre deux échelles consécutives, le changement est inférieur à 20%. Pour aller jusqu'à un facteur 2, les différentes échelles retenues ont pour valeur : 0.48, 0.58, 0.69, 0.83, 1, 1.2, 1.44, 1.73, 2.07.

L'intégration des invariants différentiels présentés à la section 4.3.1 dans un cadre multi-échelle permet ainsi d'obtenir une caractérisation robuste au groupe des similitudes.

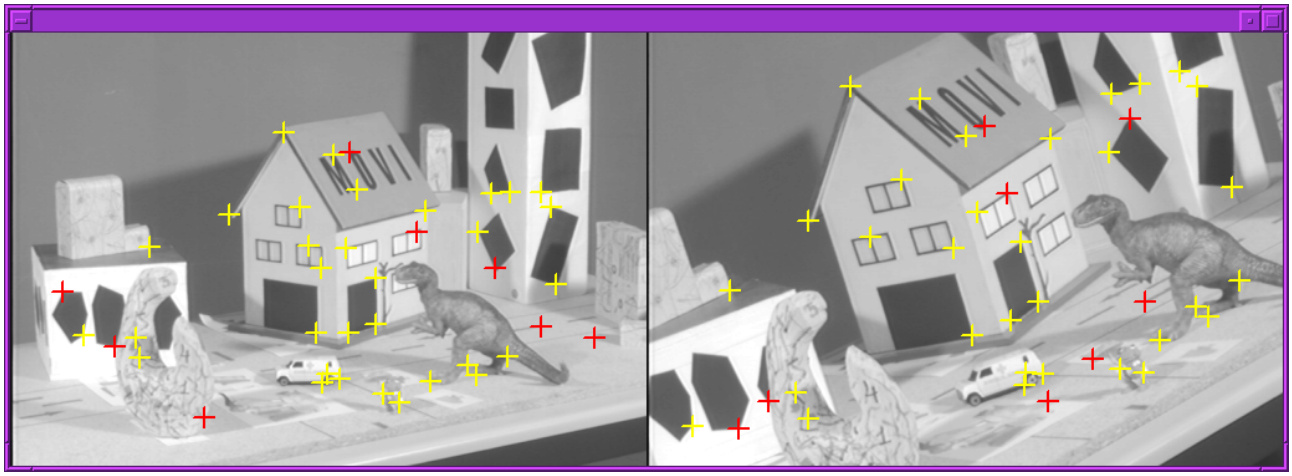

FIG. 2 — 80.0 % d'appariements corrects dans le cas d'une scène 3D

4.3.3 Robustesse à la luminosité

Il est également intéressant d'obtenir une caractérisation qui soit robuste à un changement de luminosité. Il existe plusieurs possibilités pour modéliser un changement de luminosité : translation, transformation affine, transformation monotone. Pour chacune d'elles, il est possible de calculer des invariants. Nous n'illustrerons ici que le cas de la transformation affine.

Une transformation affine de la luminance se modélise par :

$$\tilde{I}(x, y) = aI(x, y) + b$$

Une telle transformation modifie les dérivées du signal de la manière suivante : $\tilde{I}^{(n)}(x, y) = aI^{(n)}(x, y)$. N'importe quel quotient de deux dérivées est donc invariant à une transformation affine de la luminance. Dans le cas des invariants différentiels, il y a différentes manières permettant d'obtenir des invariants à une transformation affine. Une possibilité est donc de diviser les 7 dernières composantes du vecteur $\mathcal{V}$ par la puissance adéquate de la norme du gradient $(L_i L_i)^{1/2}$.

4.4 Expérimentation

Cette section présente un exemple des résultats que l'on peut obtenir. La qualité des résultats est mesurée par le pourcentage d'appariements corrects par rapport au nombre total d'appariements. Ceci a été fait sur plus de 100 paires d'images, contenant souvent plus de 100 appariements. La vérification des appariements trouvés a été faite en calculant, par une méthode de moindres carrés médians qui est robuste aux erreurs grossières, l'homographie existant entre les deux images si celles-ci représentent une scène plane, et la géométrie épipolaire dans le cas des scènes tridimensionnelles.

Il est à noter que le processus d'appariement rejette en moyenne 50% des points détectés. Ce taux de rejet est surtout dû à l'imprécision du détecteur de points d'intérêt utilisé.

La figure 2 montre les résultats obtenus sur une scène d'extérieur complexe. La transformation entre les deux images est constituée d'une rotation scène, d'une rotation image et d'un changement d'échelle. Les croix blanches indiquent les appariements corrects et les noires les appariements faux. Le pourcentage d'appariements corrects est de 80 %. Les faux appariements sont essentiellement dus aux motifs répétitifs contenus dans la scène, un coin clair ressemble en effet fort à un point coin clair d'angle similaire.

4.5 Indexation d'images

La méthode d'appariement d'images qui vient d'être présentée peut être facilement adaptée pour indexer des images. De chaque image à indexer, on extrait les points d'intérêt, que l'on caractérise comme indiqué précédemment. Cette caractérisation peut n'être faite qu'à une seule échelle pour réduire l'espace mémoire nécessaire.

Tous les vecteurs ainsi calculés sont stockés dans un arbre. À chaque niveau de l'arbre, l'espace des vecteurs est découpé selon une des composantes, de sorte que pour trouver un vecteur à un ε près dépendant de la composante, il suffise d'explorer deux cases à chaque niveau.

Lorsqu'une image qui n'appartient pas à la base doit être reconnue, ses points d'intérêt sont extraits, leur caractérisation à toutes les échelles est calculée. On compare alors ces vecteurs à ceux qui sont dans l'arbre, toute égalité (à un ε près) se traduisant pour un vote pour une image. Il ne suffit plus que de regarder quelles sont les images de la base qui ont obtenu le plus de votes.

Cette technique d'indexation a été testée sur une base contenant plus de 1000 images : images d'objets, images de tableaux de peinture, images aériennes de ville... Pour les requêtes, on a utilisé des images différentes, représentant des points de vue différents, en ne prenant éventuellement qu'une partie de ces images. Le taux de reconnaissance obtenu est supérieur à 99 %, tant que la taille de l'image requête est suffisante pour contenir plusieurs points d'intérêt. Ce taux fait de cette méthode une des plus puissantes, si ce n'est la plus puissante, à l'heure actuelle.

Un exemple est montré sur la figure 3. les deux images de gauche ont été utilisées comme requête, l'image retrouvée étant celle présentée à droite. Entre les images requêtes et l'image modèle, il y a une rotation 3D de 10 degrés de l'objet, plus dans un cas une rotation de 40 degrés dans l'image

FIG. 3 — Deux images requêtes, et l'image modèle retrouvée.

et un changement d'échelle de 1,2, et dans l'autre cas, une occultation de 50 % de l'objet.

5 Discussion

Le paragraphe 1.2 présentait les caractéristiques des techniques d'indexation. La grande majorité des propositions scientifiques que l'on trouve dans la littérature proposent d'opérer par des mesures globales stockées dans des histogrammes. Les mesures qui sont considérées sont rarement prises invariantes à des transformations, sauf un facteur d'échelle dans les intensités lumineuses pour toutes les approches qui travaillent sur des dérivées du signal. L'utilisation d'un histogramme assure une bonne robuste à des accidents comme des occultations partielles mineures. L'interface avec l'utilisateur est essentiellement réalisé à travers une sollicitation d'images pour lesquels l'utilisateur indique celle qui lui parait la plus proche de celle recherchée. C'est dans ce cadre que se positionnent les systèmes existants commercialement comme l'explique le paragraphe ci-après. Il reste cependant à imaginer des solutions vraiment opérationnelles et pour cela nous terminerons en fournissant quelques orientations de recherche.

5.1 Les systèmes commercialisés

Quelques systèmes sont opérationnels et commercialisés : VIRAGE, QBIC, EXCALIBUR, etc. Ces systèmes sont accessibles en démonstration sur le WWW. Ils opèrent tous à partir de critères globaux : l'utilisateur peut préciser s'il préfère avoir comme critère de proximité la couleur, la texture ou toute composition de ces derniers. Les détails techniques de ces systèmes ne sont pas fournis, mais il est probable qu'il s'agit de techniques basés sur des comparaison d'histogrammes. Les limites de cette façon de procéder ont été énoncées en 2. En les testant on obtient en effet des résultats pour le moins étonnants. Il n'en reste pas moins que, devant l'absence d'outils plus performants, des utilisateurs apprécient ces logiciels primitifs.

Le logiciel QBIC d'IBM offre des potentialités supplémentaires comme la recherche de images comportant des formes du type « une bouteille ». Cette fonctionnalité ne doit être considérée que sous l'angle d'une interface particulière. En effet, il n'existera pas dans un avenir prévisible de logiciel capable de détecter toutes les bouteilles dans des images et ces formes doivent donc être extraite manuellement, avec éventuellement des outils assistant l'opérateur.

5.2 Directions de recherche

Nous avons illustrés l'importance et la puissance des méthodes locales. De fait, ce genre de travaux peut facilement être étendu à d'autres familles d'invariants, par exemple en couleur. Il faut, pour cela, un minimum de compréhension de ce que sont des invariants, et surtout modéliser ce que peut être la déformation du signal qu'il faut absorber dans ces invariants. C'est un point important, mais qui n'est pas le plus difficile scientifiquement.

Plus délicat est de savoir où il faut mesurer ces invariants. [20] a proposé de prendre les points les plus caractéristiques dans la distribution des mesures locales considérées. Cette approche est séduisante car elle s'affranchit des extractions de contours et de points d'intérêt qui sont des aspects fragiles des solutions présentées en 3.1 et 4. Cette approche n'a cependant pas été validée en vraie grandeur et il reste encore à la confirmer ; la piste est en tout cas intéressante et devrait déboucher sur des solutions nouvelles et raisonnables.

Plus difficile encore est la mise en œuvre d'outils permettant une interaction réelle avec un utilisateur. Parmi les millions de caractéristiques possibles, quelles sont celles qui pourraient satisfaire la requête de l'utilisateur ? Ce problème relève a priori de l'analyse statistique. Mais nous sommes ici dans un cas défavorable où l'interaction avec l'utilisateur fournira peu de réponses, et face à cette énorme masse de données, il faudra découvrir des heuristiques performantes. Il n'est cependant pas interdit d'imaginer un système auto-adaptatif qui, mémorisant des heures de sessions, découvrirait progressivement les caractéristiques les plus performantes en fonction des contextes. Un grand chemin reste à faire.

Il est encore probablement trop tôt de parler de l'implantation informatique de ces index afin que les recherches dans de grandes bases d'images soient performantes. Il est cependant clair que les solutions offertes par les bases de données actuelles n'ont pas été conçues pour ce type d'accès.

Un grand champ scientifique s'ouvre à la communauté, alliant le traitement du signal, les outils statistiques, l'interface homme-machine et l'informatique. Ce sujet a pris une envergure considérable aux USA dans le cadre du programme « Digital Libraries ». Le démarrage en Europe est plus timide, mais bien vivant.

Références

[1] J. Ben-Arie. The probabilistic peaking effect of viewed angles and distances with application to 3D object recognition. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 12(8) :760–774, août 1990.

[2] T.O. Binford et T.S. Levitt. Quasi-invariants : Theory and exploitation. In *Proceedings of* DARPA *Image Understanding Workshop*, pages 819–829, 1993.

[3] J.B. Burns, R. Weiss et E.M. Riseman. View variation of point set and line segment features. In *Proceedings of* DARPA *Image Understanding Workshop, Pittsburgh, Pennsylvanie, USA*, pages 650–659, 1990.

[4] G. Finlayson, S. Chatterjee et B. Funt. Color angular indexing. In *Proceedings of the 4th European Conference on Computer Vision, Cambridge, Angleterre*, pages 16–27, 1996.

[5] L. Florack. *The Syntactical Structure of Scalar Images*. PhD thesis, Universiteit Utrecht, novembre 1993.

[6] J. Haffner, H.S. Sawhney, W. Equitz, M. Flickner et W. Niblack. Efficient color histogram indexing for quadratic form distance functions. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 17(7) :729–732, juillet 1995.

[7] C. Harris et M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[8] A.K. Jain et A. Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29 :1233–1244, 1996.

[9] M.S. Kankanhalli, B.M. Mehtre et J.K. Wu. Cluster based colour matching for image retrieval. *Pattern Recognition*, 29(4) :701–708, 1996.

[10] J.J. Koenderink. The structure of images. *Biological Cybernetics*, 50 :363–396, 1984.

[11] J.J. Koenderink et A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55 :367–375, 1987.

[12] B. Lamiroy et P. Gros. Rapid object indexing and recognition using enhanced geometric hashing. In *Proceedings of the 4th European Conference on Computer Vision, Cambridge, Angleterre*, volume 1, pages 59–70, avril 1996. ftp ://ftp.imag.fr/pub/MOVI/publications/Lamiroy_eccv96.ps.gz.

[13] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[14] B.M. Mehtre, M.S. Kankanhalli et A.D. Narasimhalu. Color matching for image retrieval. *Pattern Recognition Letters*, 16 :325–331, 1995.

[15] H. Murase et S.K. Nayar. Learning and recognition of 3D objects from brightness images. In *Proceedings of the AAAI Fall Symposium Series : Machine Learning in Computer Vision : What, Why, and How ?, Raleigh, Caroline du Nord, USA*, pages 25–29, octobre 1993.

[16] H. Murase et S.K. Nayar. Visual learning and recognition of 3D objects from appearance. *International Journal of Computer Vision*, 14 :5–24, 1995.

[17] Romeny. *Geometry-Driven Diffusion in Computer Vision*. Kluwer Academic Publishers, 1994.

[18] P.J. Rousseeuw et A.M. Leroy. *Robust Regression and Outlier Detection*, volume XIV of *Wiley*. J.Wiley and Sons, New York, 1987.

[19] A.H. Salden, B.M. ter Haar Romeny, L.M.J. Florack, M.A. Viergever et J.J. Koenderink. A complete and irreductible set of local orthogonally invariant feature of 2-dimensional images. In *Proceedings of the 11th International Conference on Pattern Recognition, La Hague, Pays Bas*, pages 180–184, 1992.

[20] B. Schiele et J.L. Crowley. Where to look next and what to look for. In *4th Int. Symposium on Intelligent Robotic Systems*, pages 139–146, juillet 1996.

[21] C. Schmid. *Appariement d'images par invariants locaux de niveaux de gris*. Thèse de doctorat, GRAVIR – IMAG – INRIA Rhône–Alpes, juillet 1996.

[22] C. Schmid et R. Mohr. Object recognition using local characterization and semi-local constraints. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 19(5) :530–534, mai 1997.

[23] S. Sclaroff. Deformable prototypes for encoding shape categories in image databases. *Pattern Recognition*, 30(4) :627–641, avril 1997.

[24] M.A. Stricker et M. Orengo. Similarity of color images. In SPIE *Conference on Storage and Retrieval for Image and Video Databases V*, volume SPIE-2420, pages 381–392, 1995.

[25] M.J. Swain et D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1) :11–32, 1991.

[26] C. Thorpe, M.H. Hebert, T. Kanade et S.A. Shafer. Vision and navigation for the Carnegie-Mellon Navlab. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 10(3) :362–373, 1987.

[27] B.M Romeny L.M.J Florack A.H Salden M.A Viergever. Higher order differential structure of images. *Image and Vision Computing*, 12(6) :317–325, 1994.

[28] A.P. Witkin. Scale-space filtering. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence, Karlsruhe, Allemagne*, pages 1019–1023, 1983.

[29] X. Wu et B. Bhanu. Gabor wavelets for 3D object recognition. In *Proceedings of the 5th International Conference on Computer Vision, Cambridge, Massachusetts, USA*, pages 537–542, 1995.

[30] H. Zhang, Y. Gong, C.Y. Low et S.W. Smoliar. Image retrieval based on color features : an evaluation study. In SPIE *Conference on Storage and Retrieval for Image and Video Databases*, volume SPIE-2606, pages 381–392, 1995.