



Recherche / Archives :  
numériser les images, et après ?

18 / 19 / 20 novembre 09

Research/Archives:  
Digitizing images, then what?

## INDEXATION DE MASSE ET NOUVEAUX OUTILS DE REPRESENTATION

### Indexation des images et outils de visualisation des contenus

**Marie-Luce VIAUD & Olivier BUISSON,**

*Institut national de l'Audiovisuel.*

Le sous-titre "Numériser et après" est une très bonne introduction à nos travaux, puisque nous intervenons juste après. Commençons par une présentation de l'INA. Ce qui y est intéressant, c'est l'existence de milliers de documents audiovisuels, de très nombreux utilisateurs professionnels, et des besoins qui sont très bien identifiés. À l'INA, nous voulons interpréter et analyser des contenus, faciliter l'accès aux ressources pour la notation et la consultation, donc innover en terme d'interface, mais aussi en terme d'accès. Nous utilisons toutes les modalités des documents, à savoir l'aspect image, l'aspect textuel et aussi l'aspect audio. Nous voulons aussi faciliter la réutilisation des ressources. C'est quand même un des buts de l'INA. Nous avons donc un fort besoin d'un système interactif à très grande échelle, d'aide à la notation et à la consultation. L'INA est un lieu de R & D assez unique. Nous réalisons en effet une recherche appliquée, dont vous allez voir les thématiques, nous développons, et notre objectif est de créer des prototypes applicatifs, et d'effectuer des évaluations et des tests d'usage sur les secteurs opérationnels pour voir comment nos recherches peuvent déboucher. Nous ne sommes pas censés aller jusqu'à l'industrialisation, mais nous voulons savoir si éventuellement les usages qu'on suspecte sont réels. Nous avons donc un cadre professionnel extrêmement riche et complet. Nous avons également une maîtrise complète de la chaîne de traitement. C'est assez rare puisque souvent, dans les organismes de recherche pure, ils n'ont pas de données, pas de besoins, pas de cas concret et réel, contrairement à nous. Nous disposons des données et des besoins, et nous sommes censés faire la brique scientifique au milieu, permettant d'allier ces données et ces besoins. Notre objectif dans la thématique « vie de la recherche de l'INA », c'est de décrire, chercher et visualiser, et d'interagir avec un très grand ensemble de documents. La contrainte de l'INA est qu'il faut traiter tout de suite de très grands ensembles de documents. Sinon, ce que nous faisons ne sert à rien. Le but est d'analyser, pour rechercher et accéder à des parties ou à des sous-parties d'objets multimédias. Nous allons vous faire des démonstrations pour vous montrer qu'il est possible d'accéder à des choses très précises. Nous agissons au niveau très précis du document, c'est-à-dire d'une image ou d'une vidéo. Nous souhaiterions également percevoir la richesse d'un très grand ensemble, qui satisfait une requête très précise. Nous voulons faciliter la découverte, l'utilisation et l'analyse de très grandes bases de données multimédias.

Nos hypothèses de base pour nos travaux de recherche sont de trois ordres. Nous pensons d'abord qu'il y a une inadéquation du tout automatique. En fait, qu'est-ce qui fait la différence entre une approche Web où il y a énormément de documents et une approche de type archives et

bibliothèque ? Globalement, nous essayons de maintenir une qualité et une cohérence dans tout ce qui permet d'accéder aux documents. Nous avons une certaine envie de qualité et de garantie de qualité documentaire. Or, tous les processus automatiques ne permettent pas d'atteindre cette qualité. Il y a énormément de bruits qui interviennent. Nous prôtons que le logiciel doit être au service des utilisateurs, via des processus d'interaction et d'apprentissage. C'est-à-dire que le logiciel aide, et l'utilisateur n'est pas en mode de contrôle d'un logiciel automatique qui ferait les choses à moitié. Nos hypothèses au niveau travail consistent à utiliser la représentation et la visualisation. Nous savons très bien que l'œil est très habile pour détecter des agrégats et des ensembles, et pour analyser. L'œil détecte des patterns, des différences, des densités... C'est très intéressant pour tout ce qui est l'analyse de l'information.

Une des choses typiques de l'INA, et qui est une contrainte, porte sur l'obligation de travailler à l'échelle. C'est-à-dire que nous considérons directement que nous allons travailler sur des milliards de descripteurs d'emblée. Souvent, dans un contexte scientifique, nous commençons par faire un petit truc. On se dit après que l'on va mettre plein de machines, ou alors travailler pour optimiser. Notre approche consiste à dire que nous travaillons tout de suite sur énormément de documents. Cela change du coup un peu des paradigmes.

Notre présentation va s'articuler en deux temps. Olivier Buisson va vous faire la partie "Analyse des contenus", et j'aborderai ensuite la partie "Fouille de données visuelles". Nous avons essayé d'être assez pratiques. Nous allons vous faire des démonstrations, et montrer des vidéos des prototypes que nous réalisons, pour que vous ayez vraiment une compréhension de l'utilisation, de l'utilité, et de ce que nous faisons. Je laisse la parole à Olivier Buisson.

## **Olivier BUISSON**

Je vais commencer par un exemple de ce que l'on veut faire dans notre communauté. Notre communauté a pour but de reconnaître des objets dans des images, et de les traiter, de les décrire, et de retrouver des informations dans les images. L'exemple que j'ai montré est assez caricatural, mais c'est l'objectif ultime que notre communauté veut atteindre : être capable d'interpréter une image, de donner du sens à cette image-là.

L'exemple est assez bateau. Dans notre activité, cette représentation aussi fine n'est pas notre but. Et de toute façon, il reste des dizaines et des dizaines d'années de recherche pour atteindre cet objectif-là. Nous n'en sommes qu'au début. Par contre, dans des cas particuliers, typiques de la télévision, d'archives, de ce genre de choses, il y a aujourd'hui des solutions. Nous avons des solutions pour, dans des mondes fermés comme les journaux télévisés, des flux télé, interpréter des flux télé, des images ou des sons, afin que ce soit utile pour pouvoir naviguer et rechercher efficacement. Notre but par exemple, c'est justement de dire : « Je suis rentré dans une période de JT. Cela va me permettre de changer mon processus d'indexation automatique afin de fournir des informations pertinentes aux utilisateurs par rapport à l'utilisation. »

Quel est le processus afin de pouvoir reconnaître des objets dans les images ? Nous allons traiter en fait des images, ce qu'on appelle « décrire ». Nous allons vouloir analyser les informations présentes dans les images. Il y a deux grands types de description, ce qu'on appelle les descriptions globales et les descriptions locales. En résumé, une description globale va analyser les tendances de l'image, c'est-à-dire quelles sont les principales couleurs, les principales orientations, textures, ce genre de choses permettant d'aller rechercher, dissocier ou mesurer des ressemblances entre les images. Quant aux descriptions locales, elles sont équivalentes aux premières. Nous allons essayer d'extraire les mêmes informations. Nous allons par contre nous focaliser dans des zones particulières autour d'informations saillantes, où nous allons davantage décrire des objets que des grandes tendances. À partir de ces descriptions, nous allons vouloir identifier des points communs entre des images. Le cas le plus facile à l'heure actuelle consiste à essayer de trouver des copies, comme dans le premier exemple que vous voyez avec Coluche. Il s'agit de se demander si ce sont bien des images que nous

avons réutilisées dans une production. À ce moment-là, nous allons l'utiliser pour la gestion, par exemple, de ce genre de système. Mais mettre en place des systèmes robustes a déjà demandé quelques années de recherche.

Nous avons ensuite le deuxième cas, c'est-à-dire trouver les mêmes objets. Là est le challenge. Il faut trouver des mesures de similarité et des techniques d'apprentissage qui permettent d'identifier des objets communs, afin que l'utilisateur puisse par exemple retrouver toutes les images comportant des pianos, pour l'aider à étiqueter l'existence de piano dans ces images.

Les composants utiles pour pouvoir interpréter les images sont grossièrement l'extraction d'informations dans les images, et la création de systèmes de mesure ou de similarité entre ces images-là. Comme la plupart des archives, nous manipulons des millions d'images et des centaines de milliers d'heures. Nous sommes obligés de manipuler énormément d'informations, des milliards de descriptions. Pour ce faire, nous devons développer des moteurs de recherche dédiés. Le moteur de recherche le plus connu est Google.

En fait, un travail scientifique important a démarré il y a une vingtaine d'années, pour développer des moteurs textuels, afin de structurer, naviguer et chercher efficacement dans les textes. Depuis 1990-1999, une activité a débuté pour structurer les informations visuelles et sonores de manière efficace, afin d'être capable de manipuler des descripteurs de contenus à très grande échelle. Notre spécificité à l'INA est de combiner ces trois domaines : l'extraction d'informations ; les systèmes de comparaison et de description de ces informations ; et des moteurs de recherche spécifiques, appelés moteurs de recherche vectoriels, afin de manipuler ces descriptions à grande échelle. Comme l'a signalé précédemment Marie-Luce Viaud, manipuler des corpus de très grande taille est une spécificité à l'INA. La communauté a considéré pendant longtemps que le fait de passer à l'échelle, c'est-à-dire de traiter des corpus de taille importante, était un problème industriel.

Nous avons essayé d'inverser le problème, et nous nous sommes posé la question : qu'est-ce que cela a changé dans notre activité de recherche ? Nous avons tenté de l'introduire directement dans notre problématique du point de vue théorique.

Nous avons commencé cette activité il y a neuf ans. Ces deux dernières années, la communauté commence à se poser la même question. Cette problématique est au centre de nos activités. Aujourd'hui, comme nous avons réussi à développer des méthodologies permettant l'accès à ces descriptions, nous nous interrogeons : cette échelle, cette masse de données, est-elle un inconvénient ou un avantage ? Notre hypothèse part de l'idée que c'est un avantage. Il y a des phénomènes qui apparaissent dans des corpus de très grandes tailles, des corpus audiovisuels, qui n'apparaissent pas à petite échelle. Dans notre communauté, le domaine scientifique qui permet de traiter ce type de problème s'appelle le *data meaning*. Je vous ai illustré ce phénomène avec deux cas : l'analyse de collections de programmes TV ; et l'analyse de flux TV.

Si vous prenez par exemple notre présentateur favori de France 2, quand vous regardez le journal de cette chaîne, c'est lui qui structure le journal. Quand vous analysez visuellement, un reportage arrive dès qu'il arrête de parler, et dès qu'il reprend la parole, c'est un nouveau sujet qui arrive. Nous voyons donc qu'il y a une structuration naturelle du visuel grâce à sa présence. Nous pourrions inverser le paradigme. L'animateur est l'image la plus fréquente et cette fréquence structure le journal. Mais pour le documentariste, ce n'est sûrement pas le présentateur qui est intéressant. C'est tout ce qui est reportage. Dans un journal télé, c'est la rareté qui est intéressante et précieuse. Par contre, si nous allons dans ce paradigme-là, nous nous rendons compte que le générique au sein d'un journal télé est aussi rare qu'un reportage. Le générique est présent au maximum deux fois. Il n'est pas très fréquent. Mais quand nous commençons à traiter des volumes importants, nous avons utilisé dans ce cas-là, une collection de 700 heures de journaux télé différents. À ce moment-là, le générique n'est plus rare. Il devient fréquent. Nous arrivons ainsi à faire la distinction, de manière automatique, entre tout ce qui est reportage, générique et objet récurrent, mais qui ne fait partie que de l'habillage.

Le cas du générique est important pour se localiser et aller vite aux endroits intéressants, mais n'apporte pas d'informations. Par contre, les deux autres cas que vous voyez sur la carte en bas qui

introduit des sujets, et sur l'illustration chiffrée en deuxième à droite, sont très importants pour un documentaliste. Cela permet premièrement de localiser le reportage. La deuxième représentation reprend des informations chiffrées très importantes, qui sont en liaison avec le reportage.

Grâce à une analyse à grande échelle d'une collection de ce JT, nous arrivons automatiquement à proposer à l'utilisateur des éléments structurants. Ces derniers permettent de rechercher les informations importantes. La carte permet par exemple de savoir dans quelle région se déroule le reportage, les données chiffrées qu'il va falloir récupérer, et aussi de savoir quand le reportage démarre, parce que nous avons le présentateur.

Nous allons vous faire une démonstration de l'utilisation de ces informations dans le cadre d'un Player avancé de journaux télé.

## Marie-Luce VIAUD

Ce Player avancé en est pour l'instant au tout début. C'est un dernier prototype qui vient juste d'être mis en place. Nous avons donc proposé une représentation de 100 JT, qui s'est faite totalement automatiquement. Une boucle est proposée pour chaque reportage. La ligne horizontale structurante au milieu représente le présentateur. Il y a une quinzaine de reportages dans ce JT. Nous nous apercevons que le reportage le plus long est un duplex.

Il se trouve que sur les JT que nous avons déjà annotés à l'INA, nous pouvons aussi associer les métadonnées. Sur les prototypes que nous avons à l'INA, elles le sont. Mais nous avons là un souci de session de fenêtres. Avec le Player, nous pouvons jouer la vidéo, etc. Vous voyez une fonction « search » qui est en train d'être implémentée avec des partenaires allemands, qui font de la transcription. À partir de la transcription de ce qui est dit, cela nous permettra de faire des recherches, et d'aller chercher les passages du JT qui correspondent éventuellement.

Nous pouvons donc avoir une recherche texte qui va nous repositionner par rapport aux métadonnées d'analyse faites par des documentalistes. Les modules scientifiques segmentent, mais ils n'appliquent aucune connaissance sur le JT. Or, si nous voulons avoir de la qualité, il faut appliquer de la connaissance. Et ce n'est pas l'outil qui va le faire. Typiquement, pour avoir une documentation de qualité, il y a des tâches que la machine peut faire, ce qui peut aider le documentaliste, parce que segmenter n'est pas une tâche drôle. Par contre, la connaissance que nous allons mettre dans le JT est clairement une connaissance documentaire classique. Il est aussi possible de faire des recherches dans des annotations de manière classique, ou dans une transcription du JT, en sachant que les transcriptions ne donnent pas actuellement des résultats exacts, et contiennent beaucoup d'erreurs. Cependant, cela peut se révéler utile si nous ne disposons pas d'accès. Il n'existe pas non plus d'égalité des langues par rapport à la transcription. En allemand, toutes les syllabes sont prononcées, et les transcriptions allemandes sont donc bien meilleures que celles en langue française. En langue française, ce n'est pas encore au point. Néanmoins, cela peut être intéressant pour faire des recherches. C'est toujours mieux que rien.

## Olivier BUISSON

Autre exemple que nous sommes en train de développer et expérimenter : nous avons mis en place une station qui permet de récupérer sept flux télé en parallèle, et de manière synchronisée. Ceci nous permet de les analyser et d'extraire des informations, afin de repérer rapidement les reportages, de naviguer et de rechercher efficacement. Nous avons bien représenté les différents reportages et les différents moments du journal, ce qui permet un accès rapide.

La deuxième partie que nous envisageons dans les traitements automatiques consiste à aller déterminer automatiquement dans des flux télé ce qu'il est intéressant de repérer. Il existe deux

aspects intéressants suivant l'utilisateur. Si je suis un publicitaire, ce qui m'intéresse consiste à suivre les publicités. Si je suis documentaliste, ce sera des choses que je n'ai pas documentées, qui ne sont pas de la publicité. L'important est d'avoir un système automatique qui permet de filtrer des choses qui sont intéressantes pour l'utilisateur.

Le premier exemple assez classique est de trouver les documents rares dans des flux télé, ce qui permet de déterminer tout ce qui n'est pas de l'habillage, et qui est donc utile à trouver pour les documentalistes. D'autres modèles permettent de naviguer efficacement dans les flux télé. Ce sont sur les émissions récurrentes comme la météo, ou des spots publicitaires. Ces deux derniers exemples permettent de segmenter plus de 50 % d'un flux télé. Vous avez un outil qui permet de sauter aux événements des grands changements de vos flux télé. Ceci vous permet d'avoir, dans un « visualiseur » avancé de vidéo, une méthodologie qui permet de naviguer rapidement.

Nous essayons de relever actuellement un deuxième challenge. Grâce à cette structuration de flux et cette analyse de flux, nous devons déterminer les documents importants. Nous tentons aujourd'hui de travailler sur la diffusion fréquente de reportages dans des périodes assez courtes. Cela permet par exemple de déterminer les événements importants par semaine. Nous essayons aujourd'hui dans notre corpus, comme nous savons que les événements se sont produits, de détecter par exemple automatiquement la mort de Michaël Jackson, ou la chute du mur de Berlin. Nous faisons en ce moment des tests sur quelques mois de télé.

## **Marie-Luce VIAUD**

Nous allons vous faire une démonstration du moteur de recherche multimédia VITALAS. Ce dernier est un projet européen auquel nous participons au même titre qu'INRIA, ADS, le Fraunhofer (un institut de recherche allemand), etc.

## **Olivier BUISSON**

Je vous ai expliqué précédemment ce que nous appelons un moteur de recherche vectoriel. Le but du jeu est de rechercher dans de très grandes masses de documents. Cela permet de manipuler plus de descripteurs ou plus de descriptions fines d'image. Dans notre démonstration, cette technologie est utilisée pour faire de la recherche de logos ou d'objets.

## **Marie-Luce VIAUD**

Le corpus sur lequel nous effectuons cette recherche se compose de 100 000 images de BELGA, une agence de presse belge.

Ce moteur est disponible à tester par le public pour toute personne intéressée. Nous savons que les 100 000 images de cette agence belge concernent essentiellement le sport. Nous tapons par exemple olympique, soit une recherche textuelle normale, et nous allons ensuite chercher sur une image particulière : le logo olympique. Nous faisons ensuite une recherche à partir de ce bout d'image dans l'ensemble des 100 000 images. C'est un peu long. Mais nous constatons que la première image est la même. Le logo olympique apparaît sur les tee-shirts des participants. Nous constatons que le logo olympique est présent dans beaucoup de choses. Nous voyons également que quelques ratés apparaissent, telle cette grille qui ressemble au logo. Mais globalement, cela fonctionne assez bien.

Nous allons effectuer une autre recherche pour vous montrer le fonctionnement. Il ne s'agit plus d'une similarité d'image locale, mais d'une similarité d'image globale. La tendance est globalement en rouge, et nous nous apercevons que les résultats ont aussi une tendance en rouge.

Nous allons réaliser une autre démonstration que nous avons été très contents de tester hier soir, pour vous montrer la performance de la recherche locale. Nous allons tenter le petit Marlboro à moitié caché. Nous voyons que les résultats sont assez excellents, puisque dans beaucoup d'images apparaît un petit Marlboro.

La numérisation sert à accéder aux images d'une autre façon que ce qui est possible actuellement. Ce genre de chose n'est absolument pas réalisable avec des images non digitales.

## **Olivier BUISSON**

Je précise que cette technologie a été développée en collaboration entre l'INRIA sur la partie « Description », et l'INA sur la partie « Moteur de recherche vectorielle ».

## **Marie-Luce VIAUD**

La grande échelle ne permet pas uniquement de traiter de grands ensembles d'images, mais aussi de traiter très précisément des ensembles d'images moins grands, comme le montre cette démonstration. Nous voyons donc l'étendue des possibilités, et notamment ces énormes perspectives qui s'ouvrent.

Je vais vous parler à présent de fouille de données visuelles. Il s'agit d'un grand axe que nous avons développé en nous basant sur ce que nous trouvions ennuyeux au départ. Quand on est grand lecteur, que fait-on en allant dans une grande bibliothèque ? Nous faisons la même chose qu'Umberto Eco : nous n'allons pas chercher exactement ce que nous voulons où nous voulons, sur telle étagère. Mais nous regardons. Nous fouinons jusqu'au moment où nous allons trouver quelque chose, tel un livre posé à côté d'un autre, un auteur que nous apprécions, nous permettant ainsi d'accéder à notre souhait. Nous essayons de trouver dans une bibliothèque, quand nous ignorons où trouver l'objet de nos recherches. Notre idée était de trouver des modalités d'interface qui nous permettent d'errer dans ces espaces numériques, car il n'y a pas beaucoup d'errances possibles pour l'instant.

La pertinence de la modalité visuelle. Pourquoi le visuel ? Si nous regardons les études biologiques, nous constatons que l'œil humain est extrêmement efficace pour décoder des informations. Si nous voulons avoir typiquement une notion de densité, de point de connexion, etc., nous arrivons à avoir une très rapide perception de ce genre de chose avec le visuel. La notion de « dynamisme » est également importante. À partir du moment où nous voulons représenter l'information, nous souhaitons représenter son évolution. La gêne actuelle quand nous abordons le Web est que ça bouge tout le temps : l'information arrive en flux. Nous n'avons pour l'instant à notre disposition que des recherches locales. Or, nous aimerions de temps en temps savoir ce qui se passe à une échelle plus grande, comme synthétiser les grands événements des trois derniers mois, voir comment ils ont été traités. Ils sont non seulement importants, mais comment se répercutent-ils dans l'ensemble des transmissions médiatiques. Nous voyons donc que le « dynamique » prend une importance d'autant plus grande qu'il est présent dans tout système d'information.

Pourquoi aussi la représentation visuelle ? Car elle permet de mémoriser. La spatialisation dans un plan a été utilisée de tout temps pour les cartes géographiques et pour plein d'autres applications, car nous en avons une très bonne mémorisation.

Nous voulons superviser. Nous pouvons donner les moyens de rendre ce que l'ordinateur calcule, et nous voulons allouer les moyens d'action à l'utilisateur pour agir et rentrer des connaissances dans l'ordinateur, afin de modifier son comportement en fonction de ce qu'il veut. Nous voulons faire

apprendre au système par retour d'utilisateur, ou interagir pour adapter la représentation de l'information à son usage. L'utilisateur est maître de l'action dans tous les cas.

Nous faisons des cartographies de corpus documentaires, et entre autres celui de l'offre grand public de l'INA. Cette dernière représente 20 000 heures de vidéo, et surtout 100 000 extraits et fiches documentaires. Nous ne partons pas de descriptions d'images, mais de descriptions textuelles réalisées à partir de la documentation INA. Nous allons ensuite calculer des distances entre un document et tous les autres. Notre problématique est de savoir si nous pouvons en avoir une vue générale, et gérer les proximités visuelles qui rendent compte des proximités de contenu. Dans la plupart du temps, nous avons un positionnement terrible. Tout est mélangé. Nous ne voyons rien. Notre objectif scientifique est donc d'essayer de filtrer les proximités pour ne garder que celles qui sont signifiantes, agréger les éléments les plus proches, séparer les agrégats, rendre perceptible les éléments de connexion entre des agrégats, et éventuellement faire apparaître des structures hiérarchiques et tenter de les étiqueter. L'idée est de faire émerger une structuration lisible des données si elle existe. Le but n'est évidemment pas de la créer juste pour se faire plaisir.

Vous pouvez visualiser une des premières cartes que nous avons réalisée. Elle n'est pas structurée. Vous pouvez voir les agrégats. Chaque agrégat apparaît comme un groupe de couleurs. Pour vous montrer les cohérences, j'ai mis des zooms. Nous étions ainsi du côté de *la cuisine des Mousquetaires*, qui est cet agrégat rose. Si nous lisons la carte, nous nous rendons compte que tout ce qui concerne le cinéma et la culture est groupé dans ce coin-là. Une autre partie de la carte correspond aux guerres, par exemple la guerre d'Algérie représentée par ce petit cluster violet. Toute cette partie-là concerne globalement la guerre. Celle-là est plutôt politique. Nous voyons donc qu'à partir des distances entre les documents, sans avoir aucune connaissance du contenu, nous obtenons à la fin une carte ayant « segmentiquement » du sens.

Je vais maintenant vous montrer une petite vidéo sur une autre carte. Cette dernière représente une cartographie du corpus Wikipédia Tourisme. C'est une collaboration que nous avons en projet avec Xerox. Parmi 56 000 documents, nous avons en fait extrait automatiquement les images et les textes associés des pages Wikipédia. Nous avons extrait les explications textuelles et visuelles, et nous avons cartographié le corpus à partir de ces descriptions. Tout est automatique pour l'instant sur ces corpus-là, sachant que dans les pages Wikipédia, le texte est très structuré. Nous partons de données très propres.

Je vais vous montrer une petite vidéo pour que vous ayez une idée du fonctionnement. Nous n'allons pas faire toute la vidéo, mais globalement les 56 000 images sont là. Vous retrouvez ici toutes les images naturelles, et tout le bloc qui était en bas correspond à des images synthétiques. Concernant ces dernières, nous voyons que tous les drapeaux sont bien groupés ensemble. Nous avons réussi en plus à étiqueter des sous-groupes et des sous-drapeaux. Je vais maintenant voir ce qu'est ce cluster rouge énorme : il est étiqueté « montagnes ». Il existe également un étiquetage automatique des groupes. Nous pouvons alors vérifier qu'il s'agit bien d'images de montagnes, et plus précisément des images des Alpes. Comme vous pouvez le constater si vous avez eu le temps de le lire, les mots sont inscrits sur chaque groupe. Le dernier groupe est « Vercors », « Fleur » etc. Nous voyons bien que nous sommes dans des domaines de fleurs, etc.

Dans l'assistance à la classification d'images, l'objectif est de structurer des collections d'images et des reportages, et de les annoter avec des concepts sémantiques donnés par l'utilisateur. Ce dernier est là encore au centre. L'idée est qu'il ne connaît rien ou très peu la collection. Le but est donc : de lui faire découvrir les différents groupes d'images ; de pouvoir créer et modifier à chaque instant les classes et les images se trouvant dans les classes ; et de choisir le mode de suggestion des images à classer. Nous pouvons ainsi lui présenter les images les plus ressemblantes, les plus diverses, ou celles que le système ne sait pas classer. L'objectif est de faire gagner du temps à l'utilisateur en lui faisant des propositions de classification.

Le logiciel se présente donc ainsi. Nous allons vous faire une petite démonstration, mais pas dans son intégralité. Les images en colonne sont celles qui ne sont pas classées. L'utilisateur choisit un groupe d'images, et va créer la classe. Il va mettre la première image dedans, et l'étiqueter « Désert

australien ». Les images sont décrites visuellement au départ. Aucun label n'apparaît dessus. Il va donc choisir les descripteurs qui lui semblent les plus pertinents pour son groupe d'images. C'est un peu technique. Seuls ceux qui ont fait les interfaces peuvent les utiliser. Nous vérifions cela de temps en temps. L'image est proposée automatiquement à la classe si elle va dedans. Sinon, nous allons la rejeter. Il propose là de regarder les images les plus éloignées du groupe qu'il a créé, pour essayer de trouver la diversité de son ensemble. Les images les plus lointaines, comme nous le voyons, sont relatives à la mer. Il continue, et l'image la plus loin des deux classes qu'il a créées devient un oiseau. Je voulais juste vous montrer les principes. Nous abordons donc ici le choix du parcours de mon niveau d'images.

Le système tourne. C'est donc lui qui fait les propositions, mais l'utilisateur peut toujours intervenir pour changer, si l'image ne va pas vers la classe choisie. Nous pensons que c'est presque dommage, parce que nous pourrions finalement mettre les images classe par classe. Nous avons donc une idée de toutes les images qui vont dans la classe 1, celles correspondant à la classe 2, à la classe 3. Nous pouvons alors valider par groupe d'images pour aller plus vite. Il y a potentiellement à la fin des images qui vont dans deux groupes. Le système peut donc aussi prendre en charge l'attribution d'une image dans deux groupes différents. Nous observons donc que nous avons plein de possibilités, puisque nous pouvons traiter par classe d'images. Nous pouvons détruire ou créer des classes d'images à la volée. Nous pouvons choisir deux, trois ou quatre classes pour une image. Ce système prend donc en compte beaucoup de choses, et apprend au fur et à mesure. Les images proposées tiennent compte de la nouvelle classification à l'instant « T ». Le système apprend en fonction des images déjà classées. Au fur et à mesure, il va apprendre de plus en plus finement, et fera de moins en moins d'erreurs. Nous portons au début beaucoup d'attention à la création de nos classes, à l'absence d'erreur, etc. Nous pouvons espérer que le système arrivera à classer ces images sans trop d'erreurs.

Nous réalisons également des évaluations pour connaître l'utilité potentielle de ce système.

Je voulais vous présenter également un développement exploratoire. Il s'agit d'une image de base sur laquelle nous cliquons pour en trouver d'autres plus proches.

Nous pouvons jouer également sur la modalité Texte/Image. Nous obtenons les images qui ont des textes les plus proches du texte de l'image, ou alors nous trouvons les images qui visuellement sont plus proches de l'image demandée. Nous pouvons aussi faire des modalités mixtes, et nous développons au fur et à mesure. L'idée est de savoir si une exploration de ce type-là était plus intéressante pour l'utilisateur, qu'une exploration par liste habituelle.

Nous avons fait un test d'évaluation classique avec dix utilisateurs, test auquel les documentalistes de l'INA se sont gentiment prêtés. Il en ressort qu'au niveau efficacité, il n'y a pas de grandes différences entre les deux interfaces. La nôtre n'était pas meilleure en termes de résultat, mais elle l'était par contre en termes d'efficacité. Pour trouver le même résultat, le temps était en moyenne moins long, et l'utilisateur voyait dix fois moins d'images avec cette interface qu'avec l'autre. Elle est donc moins fatigante, dans la mesure où elle demande moins de sélection à l'utilisateur.

Il nous avait été demandé de faire de la modalité et du retour de pertinence. Sur ce même type d'interface, nous avons branché un retour de pertinence consistant à dire quand des images sont pertinentes ou non pour la recherche. Le système apprend donc au fur et à mesure de la même façon. Plus nous sélectionnons d'images pertinentes, plus nous allons converger vers les images pertinentes de la base. Et comme nous donnons des exemples qui ne sont pas pertinents, nous accélérons encore cette performance. Il existe une possibilité de dialogue « utilisateur-système » qui permet d'améliorer les choses.

Nos perspectives sont : segmenter les flux télévisuels avec encore plus de contraintes de qualité et de précision ; segmenter les émissions et les reportages ; chercher des objets, et notamment des objets d'habillage, et chercher les endroits où trouver les noms des personnages. Notre but est d'accélérer l'annotation, d'accéder à ce qui est important dans les images pour l'annotateur.

Nous commençons à réaliser quelque chose de très amusant, souvent demandé à l'INA, mais sans beaucoup de moyens financiers en face. Cela consiste à faire de la recherche de personnes à partir de reconnaissance de visages et de paroles. C'est d'un projet actuellement en évaluation, et nous espérons une issue favorable. L'idée est de faire un Player vidéo enrichi où il sera possible d'ajouter au fur et à mesure, selon la demande de l'utilisateur, des options d'enrichissement d'informations.

Nous souhaitons dans un avenir plus ou moins proche corréliser les flux télévision et le Web. Nous avons la chance à l'INA d'avoir le dépôt légal du Web pour tout ce qui concerne l'audiovisuel. Nous voyons bien qu'énormément de choses sont à faire dans ce domaine, puisque des sources sont communes à ces deux canaux multimodaux. Le but est d'analyser et synthétiser des flux d'informations pour l'aide à la documentation.

## Un Intervenant dans la salle

J'aimerais savoir si vous avez eu vent d'applications concrètes de votre système ? Dans le domaine du cinéma par exemple, j'imagine que nous serions très intéressés par vos possibilités d'analyse d'images. Quand nous trouvons une bobine dont nous ne connaissons pas le contenu, votre système pourrait-il s'avérer utile ? Il ne s'agit pas ici d'images fixes, mais carrément d'une séquence. Dans ce cas-là, votre système pourrait-il fonctionner ?

## Marie-Luce VIAUD

Il faut juste numériser la bobine. Nous cherchons effectivement des séquences similaires en vidéo, mais également sur de multiples supports comme les bandes cinéma, à des définitions différentes. C'est tout à fait possible. Il faut quand même que ce soit numérisé.

## Olivier BUISSON

Des applications ont déjà été développées. Nous avons imaginé de le faire, mais cela ne s'est pas réalisé par manque de temps. Il existe déjà des applications et des travaux similaires pour le tri de rush, et je pense que cela serait utile dans le cinéma.

## Marie-Luce VIAUD

Pour la petite histoire, les réalisateurs de la série Dallas ont voulu faire une édition en haute définition, sauf qu'ils ont perdu leur montage. Ils ont donc les rushs en haute définition. Ils ont le montage en basse définition. Ils voulaient en fait un système qui puisse leur réaligner les rushs en haute définition en rushs en basse définition. Il se trouve que nous n'avons pas travaillé sur cette problématique, mais nous pouvons le faire.

## Marc VERNET

En vous écoutant, je pensais à autre chose dans ce domaine, à savoir tout ce qui est de l'ordre de la photo de cinéma. Nous avons des collections monumentales. Nous avons ici des représentants de grandes collections de photos de cinéma. Les photos sont souvent déjà numérisées. Nous pourrions appliquer des traitements de sélection de motifs par exemple, ou de scènes type, de façon tout à fait luxueuse et profitable grâce à vos systèmes.

## Marie-Luce VIAUD

C'est possible scientifiquement, et cela existe. C'est disponible à l'INA. Le problème délicat réside dans le passage du prototype de recherche au prototype industriel. Nous continuons d'avancer du point de vue scientifique, et nous pouvons fournir des choses, mais nous ne sommes pas aptes à générer un prototype industriel. Rien n'empêche de le faire technologiquement. Il ne nous manque que les moyens.

### Un Intervenant dans la salle

Nous avons vu ce que vous avez réalisé sur la reconnaissance d'images qui ne sont pas en mouvement. Travaillez-vous aussi sur la reconnaissance des séquences d'images, c'est-à-dire des images en mouvement ?

## Marie-Luce VIAUD

La démonstration que nous avons réalisée est surtout la description d'images, sans la description de mouvement. A l'INA, nous avons principalement deux types de collections, sans parler du son: les images et les vidéos. Nous avons surtout utilisé l'image pour pouvoir faire des ponts entre les deux. Nous nous servons de l'information temporelle, plus de l'enchaînement de motifs d'images. Cela se fait en deux temps, pour pouvoir par exemple retrouver de manière robuste la même vidéo pour reconstruire Dallas. C'est important, car nous avons le même décor à 200 images. L'enchaînement temporel est par contre le même. Nous utilisons l'information image, mais c'est l'enchaînement qui est temporel.

Je vais donner un exemple. On peut trouver une séquence similaire à une autre séquence. Mais si quelqu'un court, il est impossible de trouver toutes les personnes qui courent.

\*\*\*\*\*

*Suivi éditorial : Loraine Pereira – chargée de mission pour le patrimoine cinématographique / INP.*