

Une Étude sur la Récupération d'Information Multimédia : Techniques, Défis et Applications

Kushal Sangwan

20319455

DIRO, Université de Montréal

Montréal, Québec

kushal.sangwan@umontreal.ca

Résumé—La Recherche d'Information Multimédia (RIM) est un domaine en pleine expansion qui applique les techniques de recherche d'information (RI) aux contenus multimédias tels que les images, les vidéos et l'audio. Grâce aux récentes avancées en vision par ordinateur, en traitement du signal et en apprentissage profond, la RIM a connu des progrès majeurs dans l'extraction, l'analyse et l'indexation de ces données complexes et hétérogènes. Cet article présente les composants et l'architecture d'un système de RIM pour les trois types de médias : vidéo, audio et texte. Nous explorons à la fois les méthodes traditionnelles et celles basées sur l'apprentissage profond utilisées pour la récupération d'information dans ces différents types de contenu. De plus, nous discutons des nouveaux types de médias émergents et de leurs impacts sur la recherche multimédia. Enfin, nous abordons les principales techniques d'évaluation employées pour mesurer l'efficacité des systèmes de RIM.

Index Terms—RIM, apprentissage profond, indexing, recherche d'information

I. INTRODUCTION

À l'ère du numérique, les contenus multimédias — images, vidéos, sons, animations — occupent une place centrale dans notre quotidien. Leur volume ne cesse de croître, alimenté par les réseaux sociaux, les plateformes de streaming, la vidéosurveillance, les dispositifs mobiles ou encore les environnements immersifs comme le métavers. Face à cette explosion de données, la capacité à retrouver efficacement un contenu pertinent parmi des millions, voire des milliards, de fichiers devient cruciale. C'est précisément dans ce contexte que s'inscrit la **Recherche d'Information Multimédia** (*Multimedia Information Retrieval*, MIR), un domaine de plus en plus stratégique en informatique.

Contrairement aux documents textuels, qui peuvent être aisément analysés et indexés à l'aide de techniques linguistiques, les contenus multimédias posent des défis spécifiques. Une image ou une vidéo ne contient pas de mots directement exploitables pour effectuer une recherche, et dans de nombreux cas, ces contenus ne sont accompagnés d'aucune métadonnée descriptive fiable. Il devient alors nécessaire de s'appuyer sur des méthodes d'analyse automatique, permettant d'extraire des caractéristiques visuelles, sonores ou structurelles. Par exemple, une image peut être décrite selon ses couleurs dominantes, ses formes, ses textures ou les objets détectés, tandis qu'une vidéo nécessite une compréhension conjointe de la séquence temporelle et du contenu audio-visuel.

Les applications de la MIR sont nombreuses et couvrent un large éventail de domaines, du grand public aux secteurs professionnels spécialisés. Dans les systèmes de recommandation, des plateformes comme YouTube, Netflix ou Spotify utilisent la MIR pour analyser les caractéristiques audiovisuelles des contenus et suggérer des recommandations personnalisées, basées non seulement sur les préférences explicites des utilisateurs, mais aussi sur les similarités de contenu. En vidéosurveillance intelligente, la MIR permet de retrouver rapidement des séquences pertinentes dans d'immenses bases de données vidéo, en identifiant des comportements suspects, des visages connus ou des événements rares — un enjeu crucial pour la sécurité publique et la gestion urbaine.

Dans le domaine médical, la recherche d'images ou de vidéos cliniques similaires (radiographies, IRM, échographies) aide les professionnels de santé à poser des diagnostics plus précis, en comparant les cas avec des exemples précédemment annotés. Un système MIR peut, par exemple, retrouver automatiquement des images de lésions pulmonaires similaires à celles d'un nouveau patient. En éducation, les étudiants ou chercheurs peuvent accéder rapidement à des extraits de cours, des démonstrations scientifiques filmées ou des animations pédagogiques, en interrogeant parfois directement une image ou une vidéo de référence.

Dans l'industrie du divertissement, la MIR permet de rechercher des scènes spécifiques dans des films, d'identifier une musique ou un son dans une base audio, ou encore de trier automatiquement de grandes quantités de contenus générés par les utilisateurs. Les musées et les archives numériques l'utilisent également pour retrouver des œuvres d'art similaires, reconnaître des lieux ou des personnages historiques, ou encore enrichir automatiquement des collections via la reconnaissance d'objets ou de visages. Enfin, dans le commerce électronique, des systèmes de recherche par l'image permettent à un utilisateur de photographier un objet pour retrouver instantanément des produits similaires en ligne.

La recherche audio, en particulier, est un pilier important du MIR. Elle intervient dans de nombreux cas d'usage, comme la reconnaissance de la parole utilisée dans les assistants vocaux (Siri, Alexa, Google Assistant) pour transformer la voix en texte et exécuter des requêtes. On retrouve également l'indexation musicale et la reconnaissance de morceaux, avec

des applications comme Shazam qui identifient une musique à partir d'un court extrait grâce à des empreintes sonores. L'analyse d'événements sonores permet, quant à elle, de détecter des bruits spécifiques (accident, alarme, cris) dans des enregistrements, ce qui est utile pour la surveillance ou l'analyse contextuelle. Il est aussi possible de rechercher dans les podcasts en identifiant des épisodes pertinents sans métadonnées détaillées, en analysant directement le contenu sonore. Enfin, les systèmes de recommandation audio suggèrent des contenus similaires à partir des caractéristiques extraites de l'enregistrement d'entrée.

Cependant, de nombreux défis techniques persistent. D'abord, la grande variété des formats et des types de données rend difficile l'élaboration d'approches unifiées. Ensuite, la subjectivité dans l'interprétation humaine des images, vidéos ou sons complique leur traitement automatique : une même scène peut être perçue différemment selon les utilisateurs. Par ailleurs, l'ampleur des volumes de données disponibles impose des algorithmes performants, capables de fonctionner en temps réel.

Un des principaux obstacles reste le *fossé sémantique* : c'est-à-dire la distance entre les données brutes que la machine peut extraire (pixels, couleurs, sons) et les intentions ou concepts recherchés par l'utilisateur (comme « un moment romantique », « une collision de voiture » ou « une ambiance festive »). Ce décalage entre le langage machine et l'interprétation humaine est au cœur des problématiques de la MIR.

Dans le cas particulier de la recherche audio, plusieurs défis s'ajoutent, comme la variabilité du signal audio selon les conditions d'enregistrement (bruit de fond, qualité du micro), la complexité temporelle inhérente à ce type de données, et l'absence fréquente de métadonnées exploitables (titre, transcription, tags).

Contrairement aux textes, les contenus multimédias n'offrent pas toujours d'indices explicites comme des mots ou des phrases. Cela pousse la MIR à développer des approches spécifiques, souvent inspirées, mais distinctes, de celles utilisées en Recherche d'Information textuelle. Là où un moteur textuel s'appuie sur des index inversés et des modèles de correspondance linguistique, un moteur multimédia doit comparer des vecteurs de caractéristiques extraits automatiquement, selon des méthodes de représentation profondément différentes.

C'est dans ce contexte que l'intelligence artificielle, et en particulier l'apprentissage automatique et l'apprentissage profond, prend une place essentielle. Grâce aux réseaux de neurones convolutionnels (CNN) pour l'analyse d'images, ou aux architectures séquentielles pour la vidéo et l'audio, il est aujourd'hui possible d'extraire des descripteurs riches, proches du niveau sémantique, capables d'associer automatiquement un contenu visuel ou sonore à des concepts abstraits comme « danger », « joie » ou « collision ». Par exemple, des modèles préentraînés tels que CLIP (Contrastive Language-Image Pretraining) [1] ont récemment permis de lier directement du texte et de l'image dans un même espace sémantique,

autorisant des requêtes aussi naturelles que « un chat sur un skateboard » sans aucune annotation manuelle préalable.

Ainsi, la recherche d'information multimédia constitue un champ interdisciplinaire en pleine expansion, à la croisée de l'intelligence artificielle, du traitement du signal, de la fouille de données, des sciences cognitives et des techniques classiques de recherche d'information. Son objectif fondamental est de proposer à l'utilisateur une interaction plus naturelle, plus intuitive et plus efficace avec les données visuelles et sonores, dépassant la logique du simple mot-clé, pour tendre vers une véritable compréhension du contenu.

II. L'HISTOIRE DE MIR

Les débuts de la Recherche d'Information Multimédia (MIR) reposaient largement sur des algorithmes issus de la vision par ordinateur, tels que ceux présentés dans trois ouvrages de référence : Ballard et Brown [2], Levine [3] et Haralick et Shapiro [4]. Ces travaux se concentraient principalement sur la recherche de similarité à partir de caractéristiques visuelles extraites d'images, de vidéos ou d'audios. Parmi les systèmes emblématiques de cette période figurent QBIC [5] et Virage [6], qui ont marqué les années 90. Peu après, cette approche fondée sur la similarité a été adaptée aux moteurs de recherche d'images sur le Web, comme Webseek [7] et Webseer [8]. En parallèle, des efforts ont été menés pour intégrer ces techniques de recherche par le contenu au sein de bases de données commerciales via des solutions telles qu'Informix Datablades, IBM DB2 Extenders ou Oracle Cartridges [9], [10], avec l'ambition de rendre la MIR plus accessible aux applications industrielles.

Dans le domaine de la recherche vidéo, l'accent portait dans les années 90 sur la détection robuste des limites de plans. Les approches les plus courantes reposaient sur le seuillage des distances entre les histogrammes de couleurs de deux images consécutives [5]. Hanjalic et al. [11] ont proposé une méthode innovante qui ne dépendait d'aucun paramètre manuel, fournissant automatiquement un ensemble de keyframes basé sur un modèle objectif du flux d'information de la vidéo. De leur côté, Haas et al. [12] ont introduit une méthode reposant sur l'analyse du mouvement pour déterminer les limites de plans, surpassant les méthodes fondées sur les histogrammes. Leur approche permettait également une classification sémantique des scènes (zoom avant/arrière, panoramique, etc.). Un guide plus récent sur la détection des transitions vidéo est fourni par Lienhart [13].

À la fin des années 90, une prise de conscience émerge : les systèmes de recherche par similarité basés uniquement sur des caractéristiques de bas niveau ne sont ni intuitifs ni conviviaux pour le grand public. Les systèmes conçus par les chercheurs étaient, en somme, principalement utilisables par ces mêmes chercheurs. Une nouvelle direction s'imposait : créer des interfaces plus accessibles, capables d'exploiter la richesse des contenus multimédias pour des utilisateurs non-experts. Cela nécessitait une évolution vers la compréhension de la *sémantique* des requêtes, et non plus seulement le traitement des caractéristiques brutes. Ce défi, connu sous le

nom de *semantic gap*, consistait à traduire les caractéristiques calculables des contenus multimédia en concepts de haut niveau compréhensibles par les utilisateurs. Des démonstrations précoces dans cette direction ont été réalisées, notamment sur la détection de visages humains par Rowley et al. [14] ou encore Lew et Huijsmans [15].

Peut-être le premier système à réellement prendre en compte cet écart sémantique à travers l'interface de requête, l'indexation et la restitution des résultats est ImageScape [15]. Ce moteur permettait aux utilisateurs de formuler des requêtes explicites sur des objets visuels comme le ciel, les arbres ou l'eau à l'aide d'icônes positionnées dans une interface graphique. Le système opérait sur un index Web contenant plus de 10 millions d'images et de vidéos annotées par images clés.

Il est important de noter que les moteurs de recherche par similarité fondés sur les caractéristiques visuelles ou audio ont trouvé des applications concrètes dans de nombreux domaines spécifiques [16]. Cela inclut la recherche dans des bases de données de marques déposées [17], la détection de scènes visuelles ou de mouvements similaires dans la vidéo, la sélection de morceaux musicaux avec des rythmes comparables pour des DJ [18], ou encore la détection automatique de contenu pornographique [19], [20]. Ces cas d'usage s'appuient sur l'hypothèse que les caractéristiques de bas niveau – couleur, texture, spectre sonore – sont fortement corrélées à des concepts pertinents dans le contexte applicatif.

La MIR vise en somme à rendre les connaissances accessibles, peu importe leur forme ou leur support. Comme le formule le manifeste de l'ACM SIGMM [21] : « *rendre la capture, le stockage, la recherche et l'utilisation des médias numériques une occurrence quotidienne dans notre environnement informatique* ». Ce principe reste au cœur des enjeux contemporains du domaine.

Ce bref historique s'appuie notamment sur la synthèse de Lew et al. [38], qui retrace en profondeur l'évolution des systèmes et approches en MIR jusqu'au début des années 2000.

III. COMPOSANTS D'UN SYSTÈME DE MIR

Un système de Recherche d'Information Multimédia (MIR) repose sur une architecture modulaire regroupant plusieurs composants clés, chacun jouant un rôle essentiel dans le processus de recherche, d'indexation et de récupération de contenus multimédias.

A. Interface utilisateur

L'interface utilisateur (UI) constitue le point de contact entre l'utilisateur et le système MIR. Elle permet de formuler les requêtes, de visualiser les résultats, et souvent d'interagir avec les médias affichés (lecture de vidéos, zoom sur des images, navigation dans une frise temporelle, etc.). Dans un contexte de MIR, les interfaces peuvent aller bien au-delà de la simple saisie de mots-clés : il peut s'agir de requêtes par exemple, de glisser-déposer d'une image (query by example), de dessin, ou même de recherche vocale. Une bonne interface doit aussi

intégrer des mécanismes de rétroaction (relevance feedback), permettant à l'utilisateur d'affiner les résultats en fonction de ses préférences. L'ergonomie, la rapidité de réponse, la clarté de présentation des résultats et la compatibilité avec des dispositifs mobiles ou immersifs (réalité virtuelle, réalité augmentée) sont également des critères importants.

B. Moteur de recherche

Le moteur de recherche constitue le cœur du système MIR. Il est chargé d'interpréter la requête utilisateur, de consulter les index construits à partir des contenus multimédias, de calculer les scores de similarité ou de pertinence, puis de retourner une liste de résultats classés. Ce moteur repose sur des algorithmes spécialisés selon le type de média (images, vidéos, audio), intégrant des techniques de traitement du signal, de reconnaissance de formes, ou encore d'apprentissage automatique. Dans certains cas, il intègre également des modèles de recherche sémantique, capables d'inférer des intentions à partir de requêtes approximatives ou vagues. Il peut aussi fusionner plusieurs types de données (multimodales) pour enrichir les résultats, par exemple en combinant analyse visuelle et texte descriptif.

C. Module d'indexation

L'indexation est une étape cruciale dans tout système de MIR. Le module d'indexation est responsable de l'analyse automatique des données multimédias, de l'extraction de caractéristiques pertinentes, et de leur organisation sous une forme qui permet une recherche rapide et efficace. Pour les images, on extrait généralement des descripteurs de couleurs, de formes, de textures, ou encore des objets reconnus. Pour les vidéos, on extrait aussi des données temporelles, des scènes clés (keyframes), du mouvement, ou des pistes audio. Ces caractéristiques sont ensuite normalisées et structurées dans des index, souvent inspirés de ceux utilisés en recherche textuelle, mais adaptés aux particularités des médias visuels ou sonores. Ce module est également responsable de la mise à jour régulière des index pour intégrer de nouveaux contenus ou retirer ceux devenus obsolètes.

D. Base de données multimédia

La base de données constitue le réservoir dans lequel sont stockés les documents multimédias (images, vidéos, audios, métadonnées). Elle peut être centralisée ou distribuée selon la taille du système et les exigences de performance. Contrairement à une base de données classique, une base multimédia doit pouvoir stocker des objets de grande taille, non structurés, et parfois en plusieurs formats (JPEG, MP4, WAV, etc.). Elle est souvent couplée à une base de métadonnées, contenant des informations extraites automatiquement ou manuellement (dates, auteurs, descripteurs sémantiques, annotations utilisateurs). Le système doit garantir la disponibilité rapide de ces données, tout en assurant leur intégrité, leur sécurité, et leur accessibilité sur différents supports.

E. Algorithmes de similarité

Les algorithmes de similarité jouent un rôle central dans les systèmes de Recherche d'Information Multimédia (MIR). Leur objectif est de comparer une requête, qu'elle soit textuelle, visuelle ou sonore, avec les contenus présents dans la base de données, afin de déterminer ceux qui lui ressemblent le plus. Cette mesure de ressemblance permet d'établir un classement pertinent des résultats.

De manière générale, ces comparaisons reposent sur des mesures mathématiques appliquées à des vecteurs de caractéristiques extraits automatiquement à partir des contenus multimédias. Parmi les plus courantes, on retrouve :

- **La distance euclidienne** : elle mesure la distance « directe » entre deux points dans un espace de caractéristiques. Elle est simple à implémenter et couramment utilisée pour des descripteurs continus.
- **La distance de Manhattan (L1)** : elle repose sur la somme des différences absolues entre les composantes de deux vecteurs. Elle est plus robuste aux outliers dans certains cas.
- **La similarité cosinus** : elle mesure l'angle entre deux vecteurs. Cette métrique est utile lorsque l'orientation des vecteurs est plus pertinente que leur magnitude, ce qui est souvent le cas dans les systèmes de recommandation.
- **La distance de Mahalanobis** : elle prend en compte la variance et la corrélation entre les dimensions, et s'avère utile lorsque les données sont multidimensionnelles avec des distributions non uniformes.

IV. TECHNIQUES D'INDEXATION ET DE RECHERCHE

La performance d'un système de MIR repose en grande partie sur la qualité de ses techniques d'indexation et de recherche.. Étant donné la nature variée et non structurée des médias, il est essentiel de transformer ces contenus en représentations exploitables.

A. Extraction de caractéristiques

L'extraction de caractéristiques est l'une des premières étapes du processus d'indexation dans un système de Recherche d'Information Multimédia (MIR). Elle consiste à analyser les données brutes (pixels, sons, séquences vidéo, etc.) pour en extraire des descripteurs numériques représentatifs, qui serviront de base à la comparaison entre contenus.

Pour les images, les descripteurs visuels sont généralement regroupés en trois grandes catégories :

- **Descripteurs de couleur** : ils capturent la distribution des couleurs indépendamment de la structure spatiale. L'histogramme de couleur est une méthode courante, souvent exprimée dans des espaces colorimétriques comme RGB, HSV ou CIE Lab.
- **Descripteurs de forme** : ils permettent de caractériser la géométrie des objets dans l'image, via des méthodes telles que la détection de contours, les moments de Hu, les signatures de forme ou les courbes de Fourier.

— **Descripteurs de texture** : ils décrivent les motifs visuels répétitifs, la granulosité ou la régularité dans une image. Des méthodes telles que les filtres de Gabor, les transformées de Wavelet ou les Local Binary Patterns (LBP) sont souvent utilisées.

Ces caractéristiques sont ensuite transformées en vecteurs numériques pour représenter chaque image dans un espace de caractéristiques, facilitant leur comparaison lors des requêtes.

Dans le cas des vidéos, l'extraction de caractéristiques s'étend à la dimension temporelle. En plus des descripteurs visuels issus des images clés (*keyframes*), on extrait des informations liées au mouvement (mouvements de caméra, changements de scène) ainsi que des descripteurs audio (parole, musique, bruit).

Les systèmes de recherche vidéo basés sur les concepts s'appuient de plus en plus sur des modèles d'intelligence artificielle pour détecter automatiquement des objets, scènes, actions ou situations. Ces éléments sont ensuite mappés à des concepts sémantiques, à l'aide de modèles comme :

- **CNN (Convolutional Neural Networks)** pour la détection d'objets ou de scènes fixes ;
- **RNN ou Transformers vidéo** pour l'analyse d'actions dans le temps ;
- **Modèles pré-entraînés** (ResNet, VGG, YOLO, EfficientNet, CLIP) pour la reconnaissance de concepts complexes.

Ces modèles sont souvent entraînés sur de larges bases annotées (ImageNet, COCO, ActivityNet, YouTube8M) et produisent des scores de confiance pour chaque concept détecté. L'utilisateur peut alors interroger le système en langage naturel ou via des concepts, pour retrouver les segments vidéo correspondants.

En ce qui concerne l'**audio**, l'extraction de caractéristiques vise à transformer un signal sonore brut en une représentation structurée, adaptée à l'indexation et à la recherche. On distingue plusieurs familles de descripteurs :

- **Caractéristiques spectrales** : elles représentent la distribution fréquentielle du signal audio. On y retrouve les spectrogrammes, les coefficients cepstraux en fréquence de Mel (MFCC), ou encore la transformée de Fourier discrète.
- **Caractéristiques temporelles** : elles décrivent les propriétés du signal dans le domaine temporel, comme l'énergie, le taux de passage par zéro (*Zero Crossing Rate*) ou l'autocorrélation.
- **Caractéristiques de haut niveau** : il s'agit de descripteurs plus abstraits, tels que la hauteur (pitch), la tonalité, le timbre ou le rythme, utiles dans l'analyse musicale ou émotionnelle.
- **Descripteurs spécifiques à la parole** : comme les formants, qui permettent d'identifier les phonèmes dans les systèmes de reconnaissance vocale.

Les MFCC sont particulièrement utilisés pour la reconnaissance de la parole ou la classification de sons musicaux. Les spectrogrammes permettent une analyse visuelle de l'évolution fréquentielle du signal dans le temps. Enfin, les

systèmes modernes peuvent produire des embeddings audio via des réseaux de neurones, offrant une représentation dense et sémantiquement riche, adaptée à des tâches comme la détection de genre musical, l'identification d'émotion ou la recherche de sons similaires.

Toutes ces caractéristiques extraites sont ensuite représentées sous forme de vecteurs, stockés et indexés dans des bases de données, afin de permettre la comparaison efficace entre éléments lors d'une requête.

B. Indexation multimédia

L'indexation multimédia est une étape clé dans tout système de recherche d'information visuelle ou sonore. Elle permet d'organiser les contenus à partir de leurs caractéristiques extraites (visuelles, audio, temporelles ou sémantiques) dans des structures de données adaptées, assurant ainsi une recherche rapide, précise et scalable.

Historiquement, les systèmes multimédia étaient limités par les structures de fichiers classiques ou les bases de données relationnelles, qui ne permettaient pas d'indexer efficacement des contenus complexes comme les images ou les vidéos. Les fichiers étaient souvent stockés sous forme de BLOBs (Binary Large Objects), ce qui empêchait toute recherche fondée sur le contenu. Des travaux pionniers [29], [30] ont donc exploré des alternatives plus performantes, notamment les bases de données basées sur la similarité, où les objets sont indexés selon des critères perceptuels.

Contrairement à l'indexation textuelle, qui repose principalement sur des index inversés, l'indexation multimédia fait face à des défis spécifiques : la diversité des descripteurs (couleur, forme, texture, son, mouvement), la haute dimensionnalité des espaces de représentation, et la complexité des mesures de similarité. Pour y répondre, plusieurs structures ont été développées :

- **KD-Trees** : arbres binaires adaptés à la recherche dans des espaces à faible ou moyenne dimension, largement utilisés pour les vecteurs de caractéristiques visuelles. Des améliorations, telles que l'équilibrage basé sur l'entropie [31], ont été proposées pour optimiser leur efficacité.
- **R-Trees** : conçus pour gérer des objets spatiaux et utiles dans le cadre de la recherche d'images géolocalisées ou de structures de scène.
- **Locality-Sensitive Hashing (LSH)** : permet une recherche rapide d'éléments similaires dans des espaces à grande dimension, en hachant les vecteurs de manière à préserver la proximité.
- **Index visuels inversés** : inspirés des moteurs de recherche textuels, ils sont utilisés dans les représentations par "sacs de mots visuels" (Bag of Visual Words) issues de descripteurs locaux (comme SIFT ou ORB).
- **Bases vectorielles ou orientées objet** : conçues spécifiquement pour stocker et interroger efficacement des vecteurs de caractéristiques multimodales.

Des approches complémentaires ont émergé pour répondre aux besoins de l'indexation à grande échelle. Ye et Xu [32]

ont montré que la quantification vectorielle pouvait améliorer la recherche dans de larges collections. Elkwa et Kabuka [33] ont proposé une méthode à deux niveaux basée sur des signatures visuelles (propriétés d'objets + relations spatiales), permettant une amélioration significative des performances. De leur côté, Shao et al. [34] ont couplé des caractéristiques invariantes avec une indexation optimisée pour atteindre des résultats proches du temps réel.

Dans les systèmes avancés, ces structures d'indexation sont renforcées par des approches sémantiques. Par exemple, après détection d'objets, de scènes ou d'actions dans les vidéos, les concepts extraits sont intégrés dans un *index sémantique*. Chaque segment vidéo est alors annoté avec des concepts interprétables par un humain, tels que "plage", "voiture rouge", ou "explosion". Ces concepts deviennent des points d'entrée puissants pour la recherche, permettant d'exprimer des requêtes complexes, même en l'absence de métadonnées textuelles explicites.

L'intégration de *graphes de concepts* (scene graphs), qui modélisent les relations spatiales et contextuelles entre les objets, renforce encore cette approche en rendant la recherche plus fine et contextuellement pertinente.

Enfin, l'indexation haute performance dans des environnements distribués, comme les réseaux pair-à-pair (P2P), soulève d'autres défis, notamment la malédiction de la dimensionnalité et la surcharge de communication. Muller et Henrich [35] ont proposé une approche efficace basée sur des résumés de données compacts, permettant de limiter la communication entre pairs tout en maintenant la qualité des résultats de recherche.

C. Modèles de recherche

D. Modèles de recherche

Les modèles de recherche déterminent la manière dont une requête est interprétée et comparée aux contenus indexés. En Multimedia Information Retrieval (MIR), plusieurs familles de modèles sont utilisées selon le type de données, le niveau de complexité et les objectifs du système.

1. Modèles classiques de recherche:

- **Modèle booléen** : Ce modèle repose sur une logique simple de présence ou d'absence de caractéristiques spécifiques. Il permet de combiner des critères à l'aide d'opérateurs logiques (AND, OR, NOT). Bien qu'utile pour des recherches précises, il est peu adapté aux contenus multimédia, car il ne prend pas en compte les degrés de similarité.
- **Modèle vectoriel** : Les objets et les requêtes sont représentés comme des vecteurs dans un espace de caractéristiques multidimensionnel. La similarité est généralement mesurée par des distances (euclidienne, cosinus). Ce modèle est très répandu en MIR, notamment pour la recherche d'images et de vidéos.
- **Modèles probabilistes / bayésiens** : Ils estiment la pertinence d'un document en fonction de la probabilité qu'il réponde à une requête donnée. Bien établis en recherche textuelle, ces modèles ont été adaptés aux

contenus multimédia pour intégrer des incertitudes liées à l'interprétation visuelle ou sonore.

2. Modèles fondés sur l'apprentissage automatique:

- **Modèles d'apprentissage profond** : Basés sur les réseaux de neurones, ils apprennent des représentations sémantiques riches à partir de grandes quantités de données annotées. Les réseaux convolutionnels (CNN) sont utilisés pour les images, les réseaux récurrents (RNN) pour les séquences audio, et les modèles de type transformers pour les contenus multimodaux.
- **Modèles d'auto-encodage** : Ces modèles non supervisés apprennent à encoder les données dans un espace latent compressé. Utilisés pour la réduction de dimensionnalité ou l'extraction de caractéristiques, ils sont souvent intégrés dans des systèmes de recommandation ou de recherche par similarité.

3. Modèles avancés pour la recherche multimodale et contextuelle:

- **Modèles de recherche multimodale** : Ces modèles combinent plusieurs types de données (image, texte, audio, vidéo) dans un espace de représentation commun. Ils permettent de traiter des requêtes croisées, par exemple rechercher une image à partir d'une description textuelle. Les modèles d'alignement cross-modal apprennent à relier des représentations issues de modalités différentes.
- **Modèles de recherche basée sur les graphes** : Ces modèles représentent les relations entre objets multimédia sous forme de graphes de similarité ou de graphes de scène (scene graphs). Ils permettent d'effectuer des recherches contextuelles en tenant compte des relations spatiales, temporelles ou sémantiques entre les entités du contenu.

Les systèmes modernes tendent à combiner ces approches dans une vision *multimodale*, tirant parti de la complémentarité des différentes sources d'information (visuelle, textuelle, audio, etc.). Cette intégration permet d'améliorer la pertinence des résultats, notamment dans les contextes où une seule modalité ne suffit pas à caractériser le contenu recherché.

E. Requêtes multimodales par le contenu

L'un des paradigmes les plus puissants dans la recherche d'information multimédia (MIR) est la recherche par le contenu, ou Query by Example (QBE). Dans ce modèle, l'utilisateur soumet un exemple (image, extrait vidéo, ou audio) et le système utilise cet exemple pour retrouver des éléments similaires dans la base de données. Ce type de requête est particulièrement utile lorsque l'utilisateur n'est pas en mesure de décrire précisément le contenu par des mots, mais possède un exemple visuel ou sonore. Le système extrait alors les mêmes types de caractéristiques de l'exemple que celles présentes dans la base de données, et effectue une recherche par similarité.

Recherche d'images et de vidéos basée sur des descripteurs visuels :

L'un des usages les plus courants est la recherche visuelle, qui peut être effectuée à l'aide de descripteurs de couleur, de forme et de texture. Par exemple, un utilisateur peut soumettre une image de fleur rouge et demander au système de retrouver toutes les images où la couleur rouge est dominante. De même, pour la recherche par forme, l'utilisateur pourrait soumettre une silhouette humaine et chercher des images contenant des formes similaires. Pour la recherche par texture, une image de tissu en dentelle peut être utilisée pour retrouver d'autres matériaux visuellement similaires.

Dans le cas de la recherche de vidéos, cette méthode est étendue à l'analyse des segments vidéo, où des descripteurs visuels comme la couleur, la forme et la texture sont extraits de chaque image clé et utilisés pour effectuer une recherche de segments vidéo similaires. Par exemple, un utilisateur pourrait chercher une scène avec une plage en soumettant une image représentative et trouver des vidéos où des scènes similaires sont présentes, même si ces scènes ne sont pas explicitement étiquetées.

Recherche basée sur les concepts :

Une autre approche avancée dans la recherche vidéo est la recherche basée sur les concepts, ou Concept-Based Video Retrieval (CBVR). Contrairement à la recherche qui repose uniquement sur des descripteurs visuels bruts (tels que la couleur, la forme ou la texture), CBVR s'intéresse à la signification sémantique des contenus vidéo. L'idée est de permettre à l'utilisateur de formuler des requêtes basées sur des concepts compréhensibles, comme "plage", "chat", "explosion" ou "foule". Le système tente alors de retrouver des segments vidéo où ces concepts apparaissent, même en l'absence de description textuelle explicite. L'indexation des concepts peut être réalisée en analysant les objets, scènes ou actions à l'aide de modèles d'apprentissage profond.

Requêtes pour l'audio :

En plus de la recherche visuelle, la recherche audio par le contenu est également très répandue. Dans cette approche, l'utilisateur peut soumettre un extrait sonore (comme une chanson, un bruit ou un discours) et demander au système de retrouver des éléments audio similaires dans la base de données. Cette recherche se base sur l'extraction de caractéristiques acoustiques du signal audio, telles que les coefficients cepstraux en fréquence de Mel (MFCC), les spectrogrammes, ou encore des descripteurs de rythme et de ton pour les musiques.

Quelques exemples d'utilisation dans la recherche audio incluent : - **Recherche par genre musical** : L'utilisateur soumet un extrait d'un morceau de musique et demande de retrouver des morceaux du même genre en fonction de caractéristiques comme le tempo ou l'harmonie. - **Recherche par tonalité** : L'utilisateur pourrait chercher des morceaux ayant une tonalité similaire à un extrait audio donné, utile pour des applications musicales. - **Recherche par bruit spécifique** : Dans le cas d'enregistrements d'environnement ou de sons spécifiques (par exemple, le bruit de la pluie), l'utilisateur peut demander de retrouver d'autres segments audio présentant des caractéristiques acoustiques similaires.

Ces approches peuvent être combinées pour des recherches multimodales où l'utilisateur peut, par exemple, rechercher une vidéo en fonction d'un extrait audio, ou bien une image et son contenu associé à un fragment audio.

F. Feedback de pertinence

Le feedback de pertinence joue un rôle central dans les systèmes de MIR interactifs, en particulier lorsque la formulation de requêtes précises est difficile. Après une première recherche — par exemple, une image soumise comme exemple ou un extrait vidéo — l'utilisateur sélectionne manuellement les résultats jugés pertinents ou non. Le système exploite alors ces annotations pour ajuster dynamiquement la pondération des descripteurs (couleur, forme, texture, audio, etc.) ou pour reformuler la requête dans l'espace des caractéristiques.

Cette approche permet, par exemple, de réorienter la recherche d'images vers des critères plus visuels (forme, texture) ou de corriger une requête vidéo initialement trop généraliste. Le feedback peut être implémenté via des méthodes de type Rocchio, de ré-apprentissage de modèles, ou de mise à jour interactive de réseaux de neurones. Dans tous les cas, il permet une adaptation progressive au besoin d'information réel de l'utilisateur, améliorant significativement la précision dans des contextes ambigus.

G. Navigation et Résumé des Médias Multimédia

Dans un système de recherche d'information multimédia, la phase de *retrieval* ne se limite pas à retrouver des éléments pertinents : elle inclut également la manière dont ces résultats sont présentés, explorés et synthétisés. La navigation et le résumé des médias multimédia jouent ainsi un rôle central en facilitant l'accès à l'information, surtout lorsque les collections sont volumineuses ou complexes. En améliorant la lisibilité, la structuration et la représentation des résultats, ces techniques permettent à l'utilisateur de mieux comprendre et parcourir les données récupérées, tout en réduisant le temps et l'effort nécessaires à l'analyse des contenus. Elles s'intègrent donc naturellement dans la chaîne du traitement de l'information multimédia, en tant que prolongement de l'indexation et de la recherche.

Plusieurs approches innovantes ont été proposées pour améliorer la navigation dans des collections multimédia visuelles. Spierenburg et Huijsmans [22] ont développé une méthode permettant de transformer une base de données d'images en un film vidéo. Les images sont d'abord regroupées par similarité visuelle, puis ordonnées selon leur proximité inter-cluster, avant d'être converties en séquence vidéo. Cette représentation dynamique offre à l'utilisateur une vue d'ensemble rapide et intuitive du contenu de la base.

Dans le domaine du résumé vidéo, Sundaram et al. [23] ont introduit le concept de "video skim", une version condensée d'une vidéo, constituée de segments informatifs sélectionnés automatiquement. Ce résumé visuel aide l'utilisateur à saisir l'essentiel d'un contenu long, sans avoir à le parcourir en entier. Snoek et al. [24], de leur côté, ont proposé des techniques

de regroupement par catégories, permettant une navigation à la fois thématique et temporelle dans les contenus audiovisuels.

Chiu et al. [25] ont exploré l'interaction entre vidéos et environnements virtuels, en texturisant une maquette 3D de ville à l'aide d'images extraites de vidéos. Lors d'un survol aérien, les images-clés étaient placées sur les toits des bâtiments, offrant une navigation géographique intuitive à travers une base vidéo.

L'idée de transformer une séquence audiovisuelle en représentation visuelle structurée a également été exploitée dans des projets comme celui de Uchihashi et al. [26], qui ont proposé de résumer une vidéo sous forme de bande dessinée de style manga. Les images-clés sont sélectionnées et mises en page selon leur importance narrative. Tian et al. [27] ont ensuite modélisé cette tâche comme un problème d'optimisation, afin de déterminer automatiquement la disposition la plus efficace pour faciliter la navigation dans une collection d'images.

En ce qui concerne les résultats issus de moteurs de recherche d'images, Liu et al. [28] ont comparé deux méthodes de résumé visuel : le classement linéaire des résultats et le regroupement par similarité. Leur étude a montré que le regroupement d'images similaires facilitait une exploration plus naturelle et plus rapide pour l'utilisateur.

Du côté de l'audio, une fois les caractéristiques extraites, il est crucial de structurer et classifier les données pour permettre une recherche efficace. La classification audio repose sur des techniques d'intelligence artificielle pour catégoriser les sons selon des critères tels que le genre musical, la langue, ou le type de signal. On distingue notamment :

- **Classification supervisée** : Basée sur des données annotées, elle utilise des algorithmes comme les SVM, les forêts aléatoires ou les réseaux de neurones pour apprendre à associer des extraits audio à des catégories définies.
- **Classification non supervisée** : Utilisée lorsque les données ne sont pas étiquetées, elle fait appel à des techniques de clustering (comme k-means) pour regrouper automatiquement des sons similaires selon leurs propriétés acoustiques.
- **Systèmes de recommandation audio** : Ces systèmes exploitent le filtrage collaboratif ou des architectures neuronales pour suggérer à l'utilisateur des contenus proches de ceux déjà appréciés, favorisant ainsi une navigation personnalisée et exploratoire.

La recherche audio consiste ensuite à comparer une requête (texte, extrait sonore, ou profil utilisateur) aux éléments d'une base à l'aide de mesures de similarité comme la distance euclidienne, la similarité cosinus ou la distance de Hamming. Pour accélérer cette recherche, des structures d'indexation comme les arbres k-d ou les tables de hachage sont souvent utilisées, permettant une navigation rapide dans des bases de grande taille.

V. MÉTHODES DE RECHERCHE D'IMAGES

Dans la recherche d'images, plusieurs méthodes ont évolué au fil du temps, des techniques basées sur des descripteurs

locaux faits à la main, jusqu'aux approches modernes utilisant l'apprentissage profond. Les quatre méthodes suivantes représentent cette évolution.

A. SIFT + BoW

Le **SIFT** (Scale-Invariant Feature Transform) et le modèle de **BoW** (Bag of Words) ont longtemps constitué une base solide pour la recherche d'images dans les systèmes de MIR classiques, avant l'avènement des approches fondées sur l'apprentissage profond. SIFT extrait des descripteurs locaux particulièrement robustes, invariants aux changements d'échelle, de rotation, de perspective modérée et d'illumination. Ces descripteurs sont capables d'identifier de manière fiable des points d'intérêt entre différentes vues d'un même objet ou scène. Leur pouvoir discriminant permet souvent de relier un descripteur à une région bien précise, même au sein de grandes bases d'images.

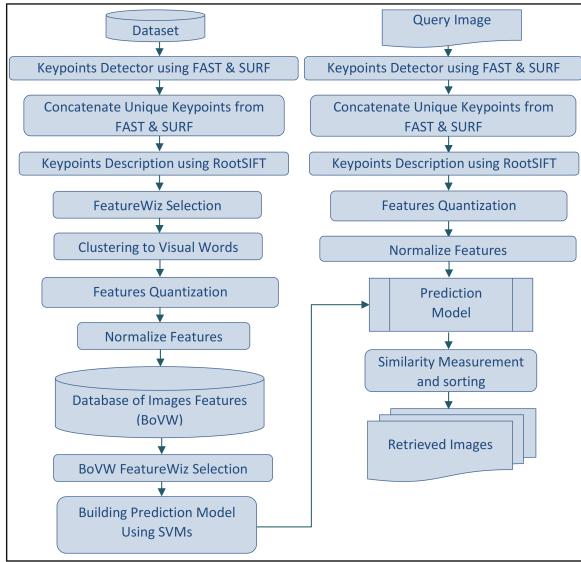


FIGURE 1. Figure extraite de [40], illustrant l'utilisation conjointe de FAST, SURF, SIFT et BoW.

Le modèle BoW repose sur la quantification de ces descripteurs en les associant à un vocabulaire visuel construit à partir d'un ensemble d'apprentissage. Chaque image est ensuite représentée par un histogramme de fréquences de mots visuels, ce qui facilite la comparaison dans un espace vectoriel simplifié. Ce tandem SIFT + BoW s'est imposé comme une référence efficace pour la recherche d'images et la reconnaissance d'objets.

Plus récemment, plusieurs travaux ont proposé des variantes ou des améliorations à cette approche. Par exemple, [40] propose une méthode hybride qui combine différents types de points-clés et techniques de sélection de caractéristiques. Elle commence par extraire séparément les points-clés FAST et SURF de chaque image, qui sont ensuite fusionnés en un vecteur de caractéristiques unifié. L'algorithme RootSIFT est utilisé pour décrire les régions autour de ces points-clés, fournissant une représentation plus robuste.

Étant donné la grande quantité de vecteurs obtenus, une sélection est effectuée à l'aide d'un algorithme de réduction de dimensions (FeatureWiz), afin de ne conserver que les plus pertinents. Ces vecteurs sont ensuite regroupés par un algorithme de clustering (k-means), formant ainsi des groupes de caractéristiques appelés "mots visuels". Ces derniers constituent le vocabulaire utilisé pour représenter chaque image sous forme d'un vecteur unidimensionnel via une quantification vectorielle.

En phase de requête, la comparaison des vecteurs d'images est d'abord une opération linéaire en $\mathcal{O}(N)$, mais elle peut être optimisée grâce à la construction d'un index inversé, réduisant la complexité à $\mathcal{O}(\log(N))$. Cette optimisation structurelle permet un passage à l'échelle plus efficace dans les bases d'images de grande taille. Figure 1 représente bien le processus avec un flowchart.

Ces approches hybrides donnent une représentation compacte et efficace des images. Toutefois, elles se heurtent encore aux limites classiques des méthodes locales, notamment leur difficulté à capturer les relations globales ou sémantiques présentes dans une scène complexe.

B. Méthodes de hachage pour l'indexation efficace d'images

Les méthodes de hachage constituent une approche incontournable pour la recherche d'images à grande échelle, en raison de leur capacité à représenter des images sous forme de codes binaires compacts. Ces représentations permettent des comparaisons rapides en utilisant la distance de Hamming, ce qui réduit significativement le coût computationnel de la recherche dans des bases de données volumineuses. Les approches de hachage peuvent être classées en deux grandes catégories : les méthodes dites classiques ou "shallow" et les méthodes basées sur l'apprentissage profond (deep hashing).

1) *Spectral Hashing : une méthode classique rapide:* Le **Spectral Hashing** (SH) [42] est l'une des premières méthodes à proposer un encodage binaire directement à partir des caractéristiques d'image tout en préservant la similarité entre les données. L'idée est de générer des fonctions de hachage analytiques en exploitant la structure spectrale (valeurs propres) du graphe de similarité des données d'entrée. Ce processus permet de transformer des descripteurs visuels, tels que SIFT ou SURF, en codes binaires de manière non supervisée. SH offre une excellente efficacité en termes de temps de recherche, car la comparaison entre images devient une simple opération binaire.

Cependant, Spectral Hashing repose entièrement sur des descripteurs conçus à la main (handcrafted features), qui sont souvent limités dans leur capacité à capturer des informations sémantiques complexes. De plus, cette méthode suppose une distribution uniforme des données dans l'espace, hypothèse rarement vérifiée en pratique. Malgré ces limitations, SH reste une solution élégante et utile pour des systèmes nécessitant une indexation très rapide avec un coût mémoire réduit.

2) *CSDH : vers un hachage sémantique avec l'apprentissage profond:* Pour pallier les limites des méthodes classiques, les approches dites de **deep hashing** ont émergé, permettant

d'apprendre simultanément des représentations discriminantes et leurs encodages binaires. Une méthode récente notable est le *Code Similarity-based Deep Hashing* (CSDH) [41], qui introduit une architecture de hachage de bout en bout basée sur un réseau convolutif profond.

Dans CSDH, une version modifiée d'AlexNet est utilisée pour extraire automatiquement des représentations sémantiques riches à partir des images. Une couche de hachage est insérée entre les couches entièrement connectées du réseau, transformant les vecteurs continus en codes binaires de longueur fixe. Chaque bit est associé à un attribut latent, et la similarité entre images est préservée lors du passage à l'espace de Hamming. Pour améliorer la robustesse de la quantification binaire, une fonction d'activation *softsign* est utilisée au sein de cette couche. Figure 2 montre bien l'architecture.

Cette méthode permet d'aligner les distances binaires avec les distances sémantiques, offrant une performance supérieure en recherche d'images tout en conservant la compacité des représentations. En exploitant l'apprentissage profond, CSDH surpassé généralement les approches comme Spectral Hashing, notamment lorsqu'il s'agit de capturer des structures sémantiques complexes dans des jeux de données d'images variées.

3) Comparaison et complémentarité: Spectral Hashing et CSDH illustrent deux approches complémentaires de l'indexation efficace : la première favorise la simplicité et la rapidité, tandis que la seconde vise une meilleure précision grâce à l'apprentissage supervisé. Dans le contexte des systèmes modernes de recherche visuelle, où les bases de données sont à la fois massives et diversifiées, les méthodes de deep hashing comme CSDH deviennent de plus en plus pertinentes, notamment dans des applications nécessitant une forte précision sémantique. Toutefois, pour des cas contraints en calcul ou sans jeu de données étiqueté, les méthodes classiques comme SH conservent leur utilité.

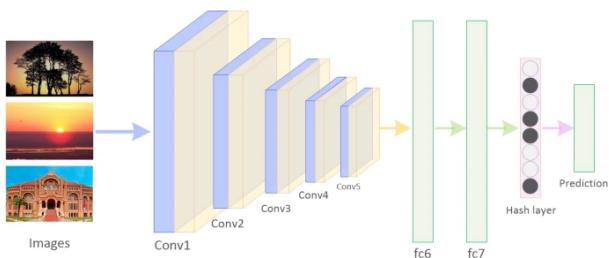


FIGURE 2. Figure extraite de [41], illustrant l'architecture de CSDH

C. Descripteurs Deep Global et agrégation régionale

Les descripteurs **Deep Global** [43] reposent sur les réseaux neuronaux convolutifs (CNN) pour extraire des représentations globales compactes des images. Ces représentations visent à capturer les informations sémantiques essentielles de l'image – telles que les objets, les scènes ou leur disposition – tout en restant efficaces pour la comparaison rapide dans un cadre de recherche d'images à grande échelle.

Une approche performante pour construire ces descripteurs consiste à utiliser les activations issues des couches convolutionnelles d'un réseau pré-entraîné (par exemple VGG16), qui fournissent des cartes de caractéristiques locales indépendantes de la taille ou du ratio de l'image. Ces cartes sont ensuite agrégées selon un schéma multi-régions, notamment le modèle dit *R-MAC* (Regional Maximum Activations of Convolutions). Ce dernier applique un *max pooling* spatial sur des régions prédéfinies réparties sur une grille multi-échelle avec recouvrement. Chaque région produit un vecteur, lequel est l2-normalisé, projeté (via une réduction de dimension de type PCA), puis l2-normalisé à nouveau. Contrairement aux méthodes classiques basées sur la concaténation, les vecteurs régionaux sont ici **sommés** pour former un seul vecteur global, garantissant ainsi une représentation de taille fixe (souvent 256 ou 512 dimensions), quel que soit le nombre de régions.

Un avantage clé de cette méthode est que toutes les opérations du pipeline sont différentiables : la sélection des régions (via ROI pooling), la projection (modélisée comme une couche entièrement connectée), la normalisation, ainsi que la somme finale. Cela permet d'intégrer cette chaîne de traitement dans une architecture CNN complète, et de l'optimiser par apprentissage.

L'apprentissage de ces représentations est généralement effectué à l'aide d'un réseau **siamois à trois branches**, exploitant une *perte triplet* pour guider l'optimisation. L'idée est de réduire la distance entre une image de requête et une image jugée similaire, tout en augmentant la distance avec une image dissemblable. Ce processus permet d'aligner la structure de l'espace de représentation avec les jugements de similarité visuelle pertinents pour la tâche de recherche.

Comme illustré dans la Fig. 3, l'architecture CNN intégrée prend en entrée des triplets d'images (requête, positive, négative) et produit pour chacune une représentation globale compacte. Un module de proposition de régions peut être intégré afin de sélectionner dynamiquement les régions les plus discriminantes pour la représentation. Lors de la phase d'inférence, chaque image est simplement traitée par le réseau pour produire un vecteur global, qui peut ensuite être comparé aux autres via un produit scalaire, ou toute autre mesure de similarité.

Cette combinaison d'agrégation régionale structurée et d'apprentissage profond end-to-end constitue l'une des avancées majeures en recherche d'images par le contenu, alliant compacité, discriminativité et rapidité d'inférence.

D. DELF : des descripteurs locaux avec attention pour la recherche d'images

La méthode **DELF** (Deep Local Features) [44] combine des descripteurs locaux extraits par un réseau de neurones convolutifs (CNN) avec un mécanisme d'attention afin de mieux capter les détails fins des images. Contrairement aux approches globales, qui se concentrent sur la représentation de l'ensemble de l'image, DELF se focalise sur des zones spécifiques et pertinentes, telles que des objets ou des points d'intérêt, ce qui permet une recherche plus ciblée et précise.

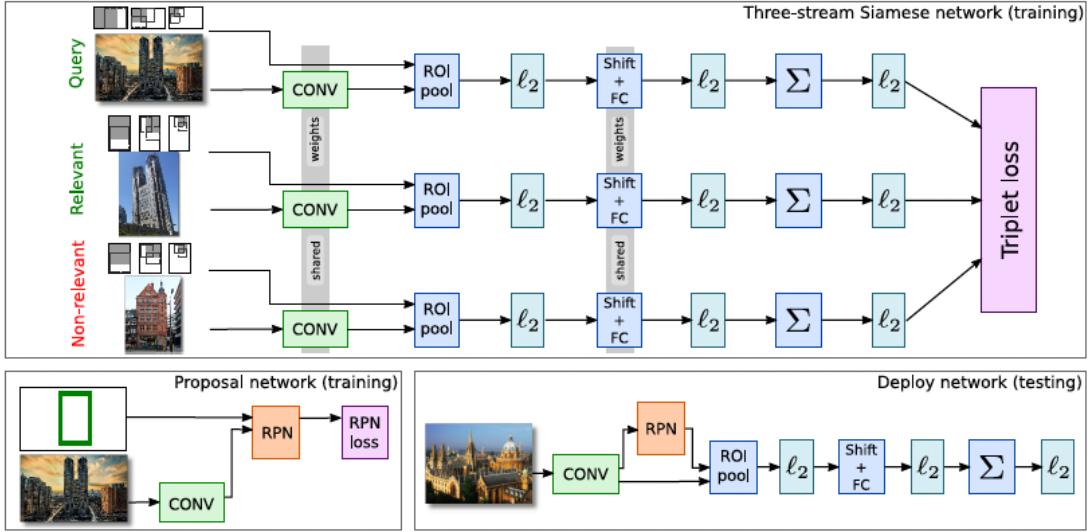


FIGURE 3. Figure extraite de [43], résumant la représentation CNN proposée.

Cette approche est particulièrement avantageuse dans des scénarios de recherche à grande échelle, tels que ceux déployés par Google, où l’efficacité et la précision sont primordiales.

comme SIFT où les points clés sont d’abord détectés, puis décrits.

Les descripteurs sélectionnés subissent ensuite une réduction de dimensionnalité par ACP (PCA) pour les ramener à 40 dimensions, avec une normalisation L2 avant et après la réduction, optimisant le compromis entre compacité et pouvoir discriminant.

Pour la recherche d’image à grande échelle, DELF utilise une indexation hybride : les descripteurs sont quantifiés via un encodage en 50 bits par Product Quantization (PQ), puis stockés dans un index inversé organisé par KD-Tree et Locally Optimized PQ. Le système permet une recherche rapide de plus d’un milliard de descripteurs en moins de 2 secondes avec un seul CPU, grâce à une stratégie d’assignation souple (soft-assignment) et à l’optimisation du nombre de régions explorées.

Lors d’une requête, chaque descripteur de l’image est comparé à ses voisins les plus proches dans l’index. Les correspondances sont agrégées par image, et une vérification géométrique avec RANSAC est appliquée pour éliminer les faux positifs. Cette étape est cruciale pour maintenir une haute précision, notamment en filtrant les correspondances issues d’images non pertinentes (distracteurs).

E. Comparaison des Méthodes

Les quatre approches présentées diffèrent significativement sur plusieurs plans, notamment le type de descripteur utilisé, la capacité de discrimination, la vitesse de recherche, ou encore l’adaptabilité aux grandes bases de données.

- **SIFT / BoW** repose sur des descripteurs locaux robustes et indépendants de l’apprentissage. Cette méthode est historiquement très utilisée dans les systèmes classiques pour sa fiabilité. Toutefois, elle présente certaines limites : elle ne capture pas les relations globales entre objets et peut être sensible à la présence d’objets vi-

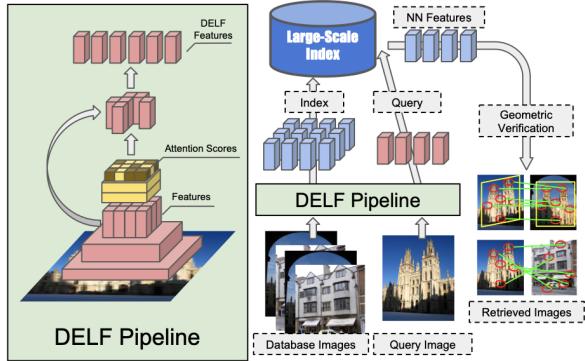


FIGURE 4. Figure extraite de [44], illustrant l’architecture de DELF

La Figure 4 illustre l’architecture générale du pipeline de traitement proposé par DELF. L’extraction des descripteurs DELF repose sur un réseau convolutionnel entièrement convolutif (FCN) dérivé de ResNet-50, appliqué à une pyramide d’images afin de gérer les variations d’échelle. Le réseau génère une grille dense de descripteurs locaux, chaque descripteur étant associé à une région locale de l’image correspondant à son champ réceptif.

Afin d’éviter l’utilisation de toutes les caractéristiques extraites — dont une grande partie pourrait être non informative —, DELF intègre un mécanisme d’attention apprenant à identifier les régions les plus discriminantes de manière faible supervisionnée. Une fonction de pondération est entraînée pour attribuer une importance à chaque descripteur, et seuls ceux avec les scores d’attention les plus élevés sont conservés pour la recherche. Cette sélection intervient après l’extraction des caractéristiques, contrairement aux méthodes classiques

Méthode	Type de descripteur	Compacité	Capacité fine-grained	Vitesse de recherche	Robustesse	Apprentissage
SIFT + BoW	Local, non appris	Moyenne	Faible	Rapide	Moyenne	Non
Spectral Hashing	Global binaire, non appris	Très haute	Faible à moyenne	Très rapide	Faible	Non
Deep Global (R-MAC)	Global, appris	Bonne	Moyenne à bonne	Rapide	Bonne	Oui
DELF (Local + Attention)	Local, appris + attention	Moyenne	Excellent	Moyenne	Très bonne	Oui (2 étapes)

TABLE I

COMPARAISON DES MÉTHODES DE RECHERCHE D'IMAGES SELON PLUSIEURS CRITÈRES.

suellement similaires ou à du bruit sémantique. L'indexation est relativement coûteuse, mais une fois les histogrammes de mots visuels construits, la recherche s'effectue rapidement.

- **Spectral Hashing** permet une recherche extrêmement rapide à grande échelle en encodant les descripteurs sous forme de codes binaires compacts. L'avantage principal réside dans la rapidité des comparaisons dans l'espace de Hamming, ce qui rend l'approche très adaptée à des bases contenant des millions d'images. En revanche, elle ne repose pas sur un apprentissage profond, ce qui limite sa capacité à capturer des relations sémantiques complexes. De plus, elle suppose une distribution uniforme des données dans l'espace de représentation, une hypothèse souvent irréaliste dans des cas concrets.
- **Deep Global**, comme l'architecture R-MAC revisitée, utilise des réseaux convolutifs profonds pour extraire une représentation globale de l'image. Ces méthodes permettent d'encoder efficacement des informations sémantiques riches grâce à des techniques d'agrégation (par somme ou max-pooling) sur plusieurs régions. Elles génèrent des vecteurs compacts et adaptés à une recherche rapide par produit scalaire. Toutefois, elles nécessitent davantage de ressources de calcul et peuvent être moins précises dans le cas d'objets petits ou partiellement visibles.
- **DELF (Deep Local Features + Attention)** combine les avantages des descripteurs locaux appris avec un mécanisme d'attention, qui permet de sélectionner dynamiquement les zones les plus pertinentes de l'image. Contrairement aux approches traditionnelles où la détection de points d'intérêt précède la description, DELF extrait d'abord les descripteurs, puis sélectionne ceux qui sont les plus informatifs. Cela lui permet d'être particulièrement efficace dans des contextes de recherche fine-grained, où la précision et la robustesse face aux variations d'apparence sont cruciales. Bien que plus coûteuse en ressources, cette méthode s'impose comme une solution de pointe pour les systèmes de recherche à grande échelle.

En résumé, l'évolution va des descripteurs locaux classiques vers des approches profondes plus sophistiquées. Le choix de la méthode dépendra fortement du contexte d'application, notamment en termes de précision requise, de taille de la base de données, et des contraintes de calcul. Nous récapitulons cette comparaison dans le tableau I, qui met en évidence les spécificités et les performances de chaque approche.

VI. MÉTHODES DE RECHERCHE VIDÉO

La recherche vidéo a évolué avec l'intégration de méthodes de plus en plus sophistiquées, des approches spatio-temporelles classiques aux modèles multimodaux modernes utilisant l'apprentissage profond. Les quatre méthodes suivantes illustrent cette progression.

A. *BoW + STIP* (Sivic & Zisserman, 2003 ; Laptev, 2005)

La combinaison de la représentation **Bag of Words** (BoW) avec les **points d'intérêt spatio-temporels** (STIP) constitue une étape fondatrice dans l'analyse vidéo fondée sur des descripteurs locaux. Deux contributions majeures, bien que publiées séparément, sont essentielles à cette approche : celle de Sivic et Zisserman [45], qui introduisent l'idée de BoW pour la recherche visuelle dans les vidéos, et celle de Laptev [46], qui généralise les points d'intérêt aux dimensions spatio-temporelles.

a) *BoW dans les vidéos* (Sivic & Zisserman, 2003).: Dans leur article *Video Google* [45], Sivic et Zisserman proposent une méthode de recherche visuelle en vidéo inspirée du modèle vectoriel de la recherche textuelle. Les étapes sont les suivantes :

- 1) **Extraction de régions locales** : les auteurs détectent des régions invariantes à l'échelle et à la rotation (affine invariant regions) dans les images clés de vidéos.
- 2) **Description locale** : chaque région est décrite à l'aide de descripteurs visuels, comme SIFT.
- 3) **Quantification vectorielle** : un vocabulaire visuel est construit par clustering (k-means) des descripteurs extraits, chaque descripteur étant ensuite assigné à un "mot visuel".
- 4) **Indexation et pondération** : les vidéos sont représentées par des histogrammes de fréquences des mots visuels, souvent pondérés par TF-IDF.

Le processus de recherche dans les vidéos passe ensuite par une série d'étapes de correspondance, qui visent à affiner la similarité entre une requête et les images indexées. La **Figure 5** illustre cette chaîne de traitement. À partir d'une région requête, des correspondances initiales sont trouvées dans les images candidates via les mots visuels. Une liste d'exclusion (*stop-list*) permet ensuite d'éliminer les mots trop fréquents, peu discriminants. Enfin, une vérification de la cohérence spatiale permet de conserver uniquement les correspondances géométriquement compatibles.

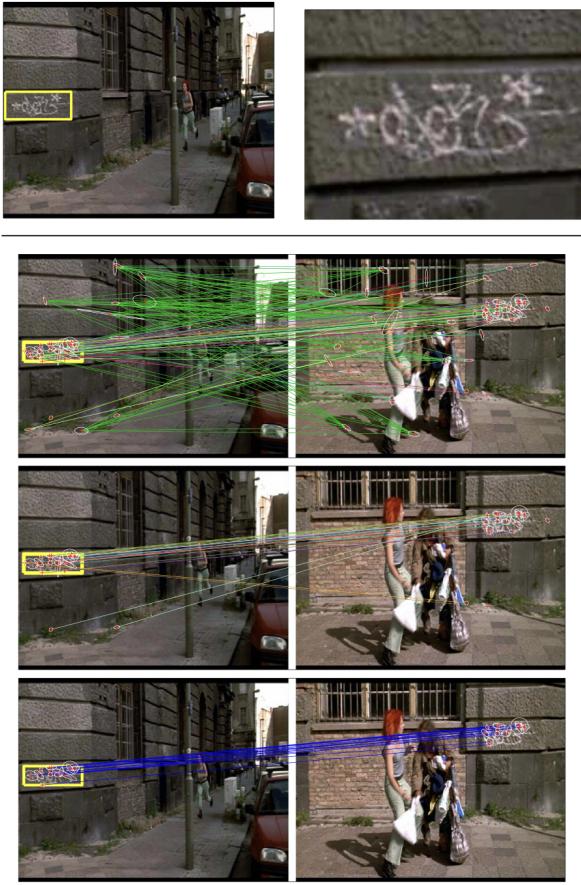


FIGURE 5. Figure extraite de [45], représentant le processus de matching des vidéos

b) *Extension temporelle* : STIP (Laptev, 2005).: Laptev propose dans [46] une extension naturelle à cette approche en introduisant des **points d'intérêt spatio-temporels** (STIP). Inspiré du détecteur de Harris, le détecteur STIP identifie des régions de l'espace-temps où les variations de luminosité sont significatives dans les trois dimensions (x, y, t). Cela permet de capturer non seulement des structures statiques (coins, bords) mais aussi des mouvements caractéristiques (sauts, gestes, collisions).

Les régions autour des STIP sont ensuite décrites à l'aide de descripteurs comme HOG (Histogram of Oriented Gradients) pour l'apparence, et HOF (Histogram of Optical Flow) pour le mouvement. Cette approche permet de capturer efficacement la dynamique locale dans les vidéos.

c) *Combinaison BoW + STIP*.: En combinant la quantification BoW avec les descripteurs extraits autour des STIP, on obtient une représentation compacte d'une vidéo sous forme d'un histogramme de mots visuels spatio-temporels. Cette approche permet de prendre en compte à la fois :

- les structures visuelles locales pertinentes (via les descripteurs),
- la localisation temporelle des événements (via les STIP),
- une représentation distribuée robuste pour l'indexation

ou la classification (via BoW).

C'est l'une des premières méthodes à intégrer l'information temporelle dans un cadre inspiré du traitement de texte, ouvrant la voie à la reconnaissance d'actions dans des scènes complexes.

B. Concept Detection (Snoek, 2006)

La méthode de **Concept Detection** se concentre sur l'identification automatique de concepts sémantiques dans les vidéos, tels que "sports", "animaux" ou "musique". Contrairement aux approches fondées uniquement sur des descripteurs visuels locaux, cette méthode introduit une couche d'interprétation plus riche en associant des motifs visuels à des catégories sémantiques compréhensibles par l'humain. Pour ce faire, elle s'appuie sur des modèles d'apprentissage statistique et des classificateurs supervisés entraînés à partir de jeux de données annotés.

Cette approche a marqué un tournant majeur dans la recherche vidéo, notamment grâce à son utilisation dans les campagnes d'évaluation comme TRECVID, où l'objectif est de détecter des concepts sémantiques à grande échelle dans des bases de données audiovisuelles complexes.

Une contribution importante dans ce domaine est celle de Snoek et al. [47], qui ont proposé une architecture modulaire appelée *semantic pathfinder*. Ce système repose sur une succession d'étapes d'analyse sémantique, organisées pour suivre un processus inverse de l'auteur (*reverse authoring process*). Chaque étape détecte des concepts de manière autonome, tout en pouvant enrichir les étapes suivantes via un modèle d'entrée-sortie standardisé.

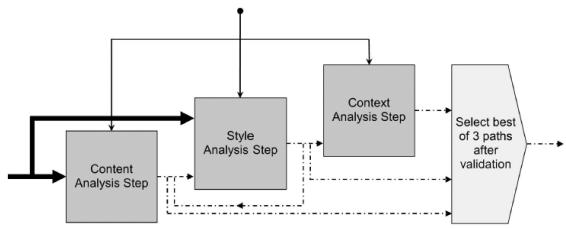


FIGURE 6. Architecture de semantic pathfinder , présentée en [47]

Le semantic pathfinder, comme nous observons dans la Figure 6, comporte trois étapes clés :

- **Content analysis** : une première analyse fondée sur les données extraites directement du contenu visuel, dans une logique *bottom-up*. Elle repose sur des descripteurs visuels et leur association avec des concepts sémantiques via des modèles statistiques.
- **Style analysis** : cette seconde étape considère la manière dont la vidéo est produite — en intégrant des éléments tels que les mouvements de caméra, les effets de montage ou le style de prise de vue — pour affiner la détection des concepts, notamment ceux liés à des structures narratives ou stylistiques.
- **Context analysis** : enfin, l'analyse de contexte cherche à détecter les relations entre concepts, ou à intégrer

des éléments textuels ou graphiques. Cette étape est particulièrement utile pour la détection de concepts complexes ou ambigus.

Chaque concept peut suivre un chemin d'analyse personnalisé à travers ces trois étapes, optimisant la précision en fonction de sa nature. Par exemple, des concepts visuels simples comme "végétation" dépendent surtout de l'analyse de contenu, tandis que des actions comme "personnes marchant" bénéficient d'une intégration plus riche des styles et du contexte. La force du semantic pathfinder réside donc dans sa capacité à adapter dynamiquement la chaîne d'analyse au concept visé.

C. Embedding Texte-Vidéo avec Données Incomplètes (Miech et al., 2018)

Cette approche introduit un modèle d'**embedding texte-vidéo** capable de gérer des données hétérogènes et incomplètes, une problématique courante dans les bases de données vidéo réelles. Le modèle proposé, appelé *Mixture-of-Embedding-Experts* (MEE) [48], apprend des représentations communes entre le texte et la vidéo en traitant simultanément des ensembles de données vidéo et image (Figure 7). Ainsi, il peut exploiter des annotations textuelles partielles ou provenant de sources variées, telles que des descriptions automatiques ou des titres de vidéos. Cette méthode permet d'améliorer la recherche vidéo en associant efficacement des informations visuelles et textuelles, même en présence de données imparfaites. L'objectif du modèle est d'apprendre un espace d'embedding commun entre des vidéos Y et des phrases textuelles X , à l'aide de fonctions d'encodage $f(X)$ et $g(Y)$ dont le produit scalaire $\langle f(X), g(Y) \rangle$ reflète la similarité sémantique entre les deux modalités. Chaque vidéo est représentée par N flux de descripteurs $\{I_i\}_{i=1}^N$, pouvant inclure le mouvement, l'apparence visuelle, l'audio ou encore les visages. Tous les types de descripteurs ne sont pas toujours présents dans chaque vidéo, ce qui reflète la nature variée et incomplète des données du web. Pour s'adapter à cette diversité, les auteurs s'inspirent de l'approche du *mixture of experts*, en apprenant un embedding spécifique pour chaque type de descripteur. Ces embeddings experts sont ensuite combinés de manière pondérée, les poids $w_i(Y)$ étant déterminés automatiquement à partir de la requête textuelle X . Ainsi, pour une phrase mentionnant une action, le modèle peut accorder plus de poids aux descripteurs de mouvement ; à l'inverse, une phrase portant sur les émotions priviliera les descripteurs de visages.

Concrètement, les descripteurs visuels de chaque flux I_i sont d'abord agrégés temporellement via un module h_i , puis passés dans un module d'embedding avec passerelle (gated embedding) g_i . Côté texte, les embeddings de mots sont également agrégés, puis encodés via un ensemble de modules f_i , chacun correspondant à un type de descripteur. Les similarités finales sont obtenues en combinant les embeddings experts selon les poids appris. Ce modèle modulaire, entraînable de bout en bout, permet une flexibilité importante et une adaptation au

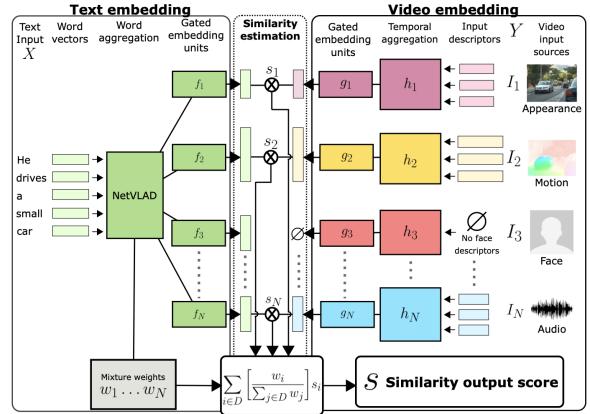


FIGURE 7. Figure extraite de [48] , illustrant l'architecture de MEEE

contenu réel des vidéos, rendant la recherche plus robuste et pertinente.

D. CLIP4Clip (Luo et al., 2021)

Le modèle **CLIP4Clip** [49] constitue une étude approfondie de l'application du modèle CLIP (Contrastive Language–Image Pretraining) à la tâche de *video clip retrieval*. L'idée est d'exploiter la puissance de CLIP, initialement conçu pour des paires image-texte, en l'adaptant aux vidéos en extrayant des frames-clés (ou clips) et en les alignant avec des requêtes textuelles. CLIP4Clip explore différentes stratégies pour agréger les similarités entre les frames vidéo et les tokens textuels (comme la moyenne, l'attention, ou les top-k similarités), afin d'identifier efficacement le clip vidéo le plus pertinent. Cette approche purement contrastive, sans supervision spécifique au domaine vidéo, démontre une performance compétitive sur plusieurs benchmarks de retrieval vidéo. CLIP4Clip illustre l'émergence de méthodes généralistes capables de tirer profit de modèles multimodaux préentraînés pour des tâches de recherche complexe.

Architecturalement, CLIP4Clip repose sur trois composants principaux : deux encodeurs mono-modaux (texte et image) et un module de calcul de similarité. La vidéo d'entrée est d'abord échantillonnée en une séquence ordonnée de frames. Chaque frame est transformée en patchs 2D et encodée via un Vision Transformer (ViT) issu du modèle CLIP (ViT-B/32). Le texte, quant à lui, est encodé à l'aide du module textuel de CLIP. La similarité est ensuite calculée à l'aide de trois stratégies étudiées : une approche sans paramètre (par moyenne simple), une approche séquentielle (qui prend en compte l'ordre des frames), et une approche "serrée" (tight) qui maximise les correspondances locales précises. Figure 8 illustre ces stratégies

Contrairement à des travaux concurrents comme celui de Portillo-Quintero et al. (2021), qui utilisent directement CLIP pour des prédictions zero-shot sans modification, CLIP4Clip propose une adaptation spécifique des mécanismes de similarité, et un entraînement de bout en bout. En plus de l'exploration des trois stratégies de similarité, les auteurs post-

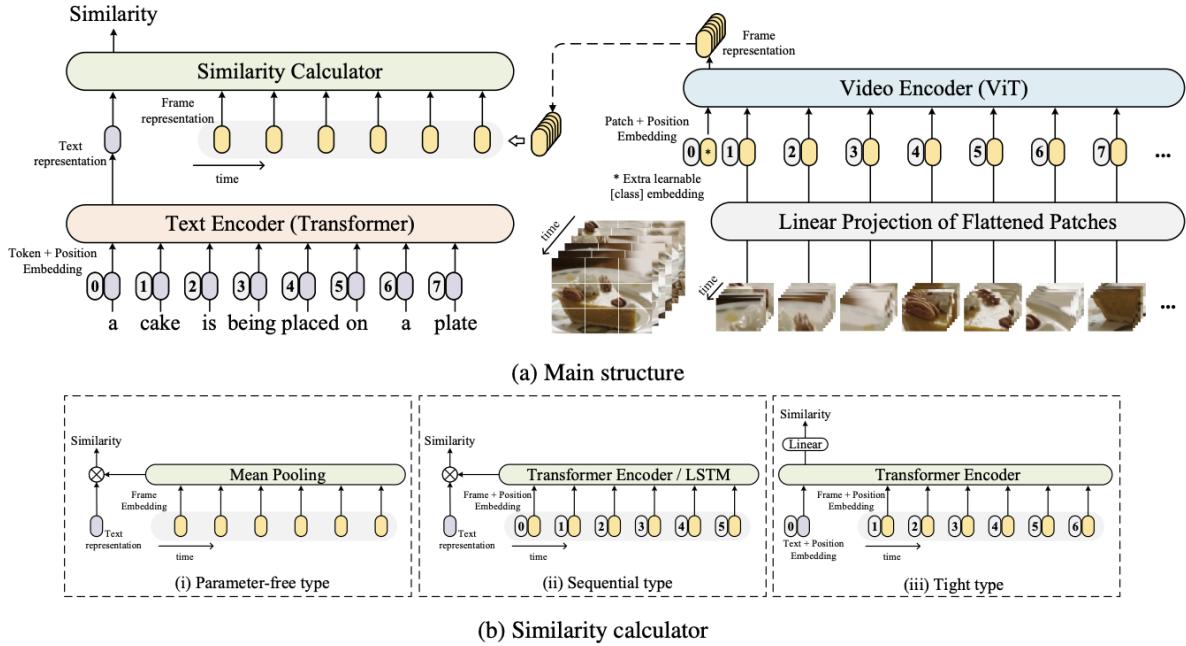


FIGURE 8. Figure extraite de [49], illustrant l'architecture de Clip4Clip

entraînent leur modèle sur un grand corpus bruité de vidéos et de textes, ce qui améliore significativement la qualité des embeddings pour la recherche.

Les résultats expérimentaux montrent que CLIP4Clip établit de nouveaux états de l'art sur plusieurs benchmarks de référence (MSR-VTT, MSVC, LSMDC, ActivityNet et DiDeMo). Plusieurs enseignements ressortent des expériences :

- Une seule image ne suffit pas à représenter correctement une vidéo pour la tâche de retrieval ;
- Le post-entraînement sur un corpus vidéo-texte à grande échelle améliore nettement les performances, notamment en zero-shot ;
- Pour les petits jeux de données, une simple moyenne des embeddings sans nouveaux paramètres est préférable ; pour les grands ensembles, l'ajout de couches comme le self-attention améliore la modélisation temporelle ;
- Une étude approfondie des hyperparamètres est essentielle pour optimiser les résultats.

E. Comparaison des Méthodes

En comparant ces quatre méthodes, plusieurs tendances et évolutions significatives apparaissent dans le domaine de la recherche d'information vidéo :

- **BoW + STIP** : Il s'agit d'une méthode pionnière reposant sur des descripteurs spatio-temporels locaux et un modèle de sac de mots visuels. Elle capte efficacement les mouvements et les objets présents dans les vidéos grâce à l'analyse locale, mais elle reste limitée par l'absence de prise en compte des relations sémantiques et contextuelles plus complexes. C'est une approche fon-

dationnelle mais aujourd'hui dépassée pour les scénarios à forte variabilité visuelle ou sémantique.

- **Concept Detection** : Cette approche marque une première transition vers une représentation sémantique du contenu vidéo. En s'appuyant sur des classificateurs pour détecter des concepts comme "musique" ou "sport", elle introduit une couche d'abstraction utile pour la recherche orientée utilisateur. Cependant, elle dépend fortement d'un vocabulaire préétabli et d'annotations de qualité, ce qui la rend moins flexible face à la diversité des contenus web.
- **Learning a Text-Video Embedding from Incomplete and Heterogeneous Data** : Contrairement aux méthodes précédentes, cette approche ne repose pas sur des descripteurs figés ou un vocabulaire fixe, mais apprend un espace partagé entre vidéos et textes, même en présence de données incomplètes. Cela permet une recherche plus robuste et plus générale dans des corpus hétérogènes. En comparaison, elle surpasse les approches classiques en adaptabilité, tout en réduisant la dépendance aux annotations manuelles.
- **CLIP4Clip** : Cette méthode représente l'état de l'art en retrieval vidéo-texte, notamment grâce à sa capacité à s'adapter à différents contextes via plusieurs stratégies de calcul de similarité. Comparée aux approches supervisées traditionnelles, CLIP4Clip se distingue par ses performances en zero-shot et sa généralisation remarquable à de nouveaux domaines, sans nécessiter de fine-tuning spécifique.

En résumé, ces méthodes illustrent clairement l'évolution du domaine : des modèles basés sur des représentations visuelles

locales et statiques vers des systèmes capables d'intégrer de l'information multimodale, incomplète et contextuelle. L'intégration du texte, de la temporalité et de la variabilité des sources devient aujourd'hui essentielle pour faire face à la richesse et à la complexité des contenus vidéo modernes.

VII. MÉTHODES DE RECHERCHE AUDIO

La recherche audio, tout comme la recherche d'images et de vidéos, a évolué au fil du temps, passant des méthodes de bas niveau basées sur des caractéristiques simples aux approches multimodales plus avancées. Les quatre méthodes suivantes illustrent cette évolution.

A. QbH / DTW (*Query by Humming / Dynamic Time Warping*)

La méthode **QbH / DTW** (*Query by Humming / Dynamic Time Warping*) [50] a été l'une des premières à proposer une approche de récupération musicale basée sur l'audio. Cette méthode utilise le *Dynamic Time Warping* (DTW), une technique permettant de comparer deux séries temporelles, pour mesurer la similarité entre un extrait chanté ou joué par l'utilisateur et une base de données d'extraits musicaux.

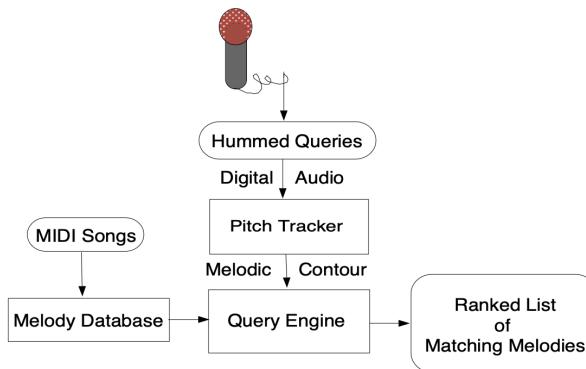


FIGURE 9. Architecture présentée en [50]

La Figure 9 illustre que l'architecture d'un système typique de QbH repose sur trois modules principaux : un module de suivi de hauteur (pitch-tracking), une base de données mélodique, et un moteur de requête. Lorsqu'un utilisateur fredonne une mélodie dans un microphone, celle-ci est numérisée puis transformée en une représentation de contour mélodique, c'est-à-dire une suite de variations relatives de hauteurs entre les notes successives. Cette représentation est ensuite comparée à la base de données à l'aide du moteur de requête, qui retourne une liste classée de mélodies correspondantes.

Cette approche s'appuie sur l'observation que le contour mélodique – la séquence des différences relatives de hauteur – est une caractéristique perceptive majeure que les auditeurs utilisent pour identifier les similitudes entre mélodies (Handel, 1989). Bien que cette méthode ait été pionnière dans le domaine de la recherche audio, elle reste pertinente aujourd'hui pour les tâches de récupération musicale symbolique et les systèmes interactifs centrés sur la mélodie.

B. Hachage pour la recherche audio (Yu, 2006)

L'approche par **hachage pour la recherche audio** [51] repose sur des techniques de réduction de dimensionnalité et d'indexation efficace des représentations audio pour permettre une recherche rapide par similarité. Contrairement aux approches d'apprentissage profond, cette méthode se concentre sur des techniques issues de l'informatique traditionnelle, notamment le hachage binaire ou local, pour transformer les descripteurs audio (souvent spectraux ou temporels) en codes compacts facilement comparables. Ce paradigme est particulièrement adapté aux bases de données de grande taille, avec des besoins de rapidité et d'efficacité computationnelle. Bien que dépassée en précision par les approches modernes, cette méthode reste représentative d'une phase intermédiaire entre les techniques purement statistiques et les réseaux neuronaux profonds.

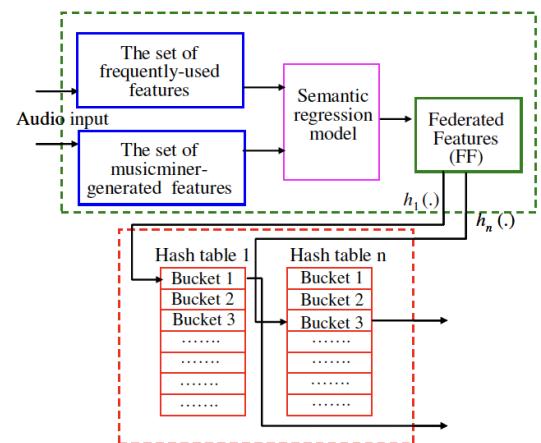


FIGURE 10. Architecture présentée en [51]

Plus spécifiquement, l'article de Yu (2006) propose un cadre de *semantic indexing* pour la recherche de variantes musicales d'un même morceau (ex. reprises chantées par différents artistes). La méthode repose sur une représentation audio sémantique appelée *Federated Features* (FF), qui résume les séquences de caractéristiques audio pertinentes. Ces représentations sont ensuite associées à des valeurs de hachage via des méthodes telles que LSH, E2LSH ou utilisées directement dans un cadre KNN. Cette combinaison permet un repérage efficace des morceaux similaires au sein de grandes bases musicales, avec une probabilité élevée de détection lorsqu'un hash est partagé entre deux pistes. Nous montrons l'architecture complète dans la Figure 10

C. Siamese Networks pour la recherche audio (Manocha, 2018)

La méthode **Siamese Networks pour la recherche audio** [52] représente une transition vers des modèles d'apprentissage automatique plus avancés dans le domaine de la recherche audio. L'approche repose sur un réseau siamois entraîné pour apprendre des représentations vectorielles de segments audio

de manière à capturer leur similarité sémantique. Chaque paire d'extraits audio est encodée dans un espace latent où la distance (euclidienne ou cosinus) reflète leur proximité perceptuelle. Ce système permet ainsi de retrouver des enregistrements similaires à partir d'un extrait de requête.

L'idée derrière cette approche est de dépasser les limitations des méthodes de fingerprinting traditionnelles, qui ne sont efficaces que pour des correspondances exactes d'événements sonores. Tandis que le fingerprinting se concentre sur des caractéristiques locales spécifiques à un enregistrement audio, le réseau siamois apprend des représentations plus générales qui peuvent capturer des similarités sémantiques entre différents événements sonores, même si ces derniers ne sont pas identiques à l'extrait de requête.

Ce système repose sur un apprentissage supervisé et utilise un modèle de réseau siamois composé de deux réseaux jumelés pour comparer les entrées audio et déterminer leur similarité (Figure 11). Concrètement, chaque réseau jumeau traite une des deux entrées audio, et le modèle est entraîné pour prédire une similarité de 1 si les extraits sont semblables, et 0 sinon. Une fois le réseau entraîné, on utilise l'un des jumeaux comme extracteur de caractéristiques pour encoder les enregistrements de la base de données. L'extrait audio de requête est lui aussi encodé via ce réseau, puis comparé à chacun des enregistrements encodés de la base afin de les classer par ordre décroissant de similarité, selon une mesure de distance — dans ce travail, la similarité cosinus et la distance euclidienne sont utilisées.

Les réseaux siamois offrent plusieurs avantages. Tous les sous-réseaux partagent les mêmes poids, ce qui réduit le nombre de paramètres à entraîner, limite les besoins en données d'apprentissage, et diminue le risque de surapprentissage. De plus, les vecteurs de représentation produits par chaque sous-réseau partagent la même sémantique, ce qui facilite grandement la comparaison entre les différentes entrées audio.

2. PROPOSED APPROACH

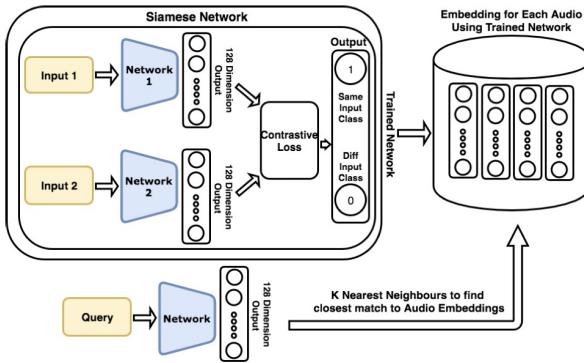


FIGURE 11. Architecture présentée en [52]

D. Réseaux à co-attention pour la recherche audio-langage (Sun, 2024)

Les **réseaux à co-attention pour la recherche audio-langage** proposés par Sun et al. [53] constituent une avancée majeure dans le domaine de la recherche audio-textuelle. Contrairement aux approches contrastives classiques qui encodent indépendamment les deux modalités avant de mesurer leur similarité, cette méthode intègre activement les représentations audio et textuelles via un module de **co-attention bilatérale**, permettant une interaction dynamique entre chaque segment audio et chaque mot de la requête.

a) *Architecture détaillée du modèle.*: Le modèle est composé de trois modules principaux :

- **Un encodeur audio** (comme CLAP), basé sur un CNN suivi d'un Transformer, qui transforme un extrait audio en une séquence d'embeddings temporels.
- **Un encodeur textuel**, basé sur un Transformer pré-entraîné (comme RoBERTa), qui encode une requête en langage naturel sous forme de vecteurs contextuels pour chaque mot.
- **Un module de co-attention symétrique**, cœur de l'architecture, qui met en relation chaque jeton textuel avec chaque frame audio, et inversement, permettant une attention bidirectionnelle.

Concrètement, le module de co-attention utilise des mécanismes d'attention croisée pour calculer :

$$T' = \text{Attention}(T, A)$$

$$A' = \text{Attention}(A, T)$$

où T est la séquence textuelle et A la séquence audio. Ces représentations croisées sont ensuite concaténées ou fusionnées via des couches résiduelles et feedforward pour former une représentation conjointe audio-texte.

b) *Empilement et itération des blocs de co-attention.*: Deux variantes de l'architecture ont été explorées :

- **Le modèle par empilement (stacked)** : plusieurs blocs de co-attention sont empilés séquentiellement, permettant un enrichissement progressif des interactions entre les modalités.
- **Le modèle par itération (iterative)** : les blocs de co-attention sont appliqués de façon récurrente, en réinjectant les sorties dans les encodeurs audio et textuel à chaque étape pour affiner dynamiquement les représentations.

Ce mécanisme itératif permet au modèle de réviser progressivement ses représentations, renforçant la correspondance sémantique entre segments audio et tokens textuels.

c) *Résultats et performance.*: Sur deux jeux de données publics — **Clotho** et **AudioCaps** — les deux variantes du modèle surpassent les approches existantes en termes de précision moyenne (mAP), avec une amélioration notable pour le modèle itératif :

- +16,6 % de mAP sur *Clotho*,
- +15,1 % de mAP sur *AudioCaps*.

Cette performance est attribuée à la capacité du modèle à apprendre des alignements hiérarchiques complexes entre les

structures linguistiques (syntaxe, dépendances lexicales) et les dynamiques temporelles de l'audio.

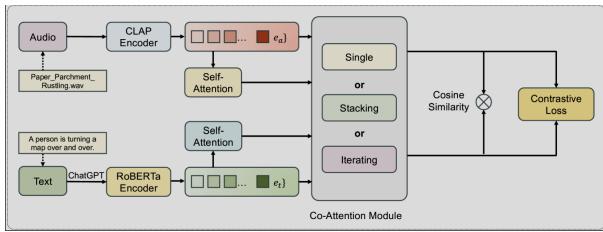


FIGURE 12. Architecture présentée en [53]

E. Comparaison des Méthodes

Voici un résumé des forces et des limites des quatre méthodes :

- **QbH / DTW** : Cette méthode est l'une des premières à traiter la recherche musicale en utilisant des techniques de comparaison temporelle comme le *Dynamic Time Warping*. Bien qu'elle ait ouvert la voie à la récupération musicale, elle est limitée dans la gestion de requêtes complexes et dans la gestion de signaux audio plus complexes.
- **Siamese Networks pour la recherche audio** : Ce modèle représente une avancée majeure dans la recherche audio, en utilisant des réseaux siamois pour encoder des extraits audio sous forme de vecteurs. Ces vecteurs permettent de mesurer la similarité sémantique entre différents extraits audio. Bien qu'efficace pour les requêtes basées sur des événements audio similaires, cette approche reste limitée par le besoin d'un apprentissage supervisé et des coûts computationnels élevés.
- **Hachage pour la recherche audio** : Cette méthode repose sur des techniques de réduction de dimensionnalité et d'indexation efficace, permettant une recherche rapide par similarité dans des bases de données audio volumineuses. Bien que dépassée en précision par les méthodes modernes, elle est encore utile pour des systèmes nécessitant une efficacité computationnelle et une rapidité de traitement.
- **Réseaux à co-attention pour la recherche audio-langage** : Cette approche représente l'état de l'art dans la recherche audio multimodale, combinant audio et texte au sein d'un même réseau. L'architecture à co-attention permet une interaction fine entre les deux modalités, capturant des correspondances complexes entre le texte et l'audio. Bien que cette méthode soit très performante pour des tâches de recherche avancées, elle nécessite de grandes quantités de données et des ressources computationnelles considérables.

Ces méthodes illustrent l'évolution de la recherche audio, passant des techniques traditionnelles basées sur des descripteurs acoustiques simples à des systèmes multimodaux modernes qui intègrent des informations audio et textuelles pour des recherches plus avancées et puissantes.

VIII. DES NOUVEAUX TYPES ÉMERGENTS DE MULTIMÉDIA

Avec l'évolution rapide des technologies numériques, de nouveaux types de multimédia émergent, permettant de capturer, d'analyser et d'interagir avec des informations sous des formes variées et plus immersives. Ces nouveaux formats repoussent les limites des formes traditionnelles de multimédia, comme les vidéos, les images et l'audio, en introduisant des concepts qui permettent de mieux refléter les expériences interactives et les environnements virtuels. Parmi ces nouveaux formats, les Meta-verse Recordings (MVR) occupent une place centrale.

A. Meta-verse Recordings (MVR)

Les **Meta-verse Recordings (MVR)**, ou Enregistrements du Métavers, représentent un nouveau type de contenu multimédia, visant à capturer une session utilisateur dans un environnement virtuel immersif et dynamique. Contrairement aux vidéos classiques qui se limitent à des images linéaires, les MVR encapsulent des données temporelles complexes issues du rendu graphique, des périphériques d'entrée/sortie et de l'environnement interactif 3D.

Steinert et al. ont exploré ces problématiques dans deux contributions récentes. Le premier travail, Information Need in Metaverse Recordings – A Field Study [54], étudie les besoins informationnels des utilisateurs et leurs comportements de recherche dans les MVR. L'étude met en évidence les spécificités des MVR par rapport aux contenus classiques (voir figure 13), notamment l'importance des données de séries temporelles liées à l'interaction et à la perception immersive. Ces résultats offrent des pistes pour concevoir des systèmes de recherche adaptés à ces nouveaux formats.

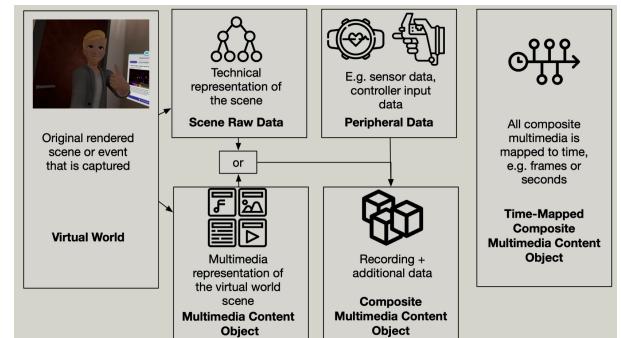


FIGURE 13. Les différents composants de MVR présentés en [54]

Dans un second article, Integration of Metaverse Recordings in Multimedia Information Retrieval [36], les auteurs analysent les défis techniques liés à l'intégration des MVR dans les systèmes MIR existants. Bien que certains algorithmes de traitement d'image et de vidéo puissent être adaptés, le papier souligne la nécessité de modules spécifiques pour interpréter les graphes de scène, les flux sensoriels ou les métadonnées comportementales générées en temps réel.

Les MVR peuvent prendre différentes formes :

- **Vidéos des médias perçus** : Elles reflètent ce que l'utilisateur voit dans le métavers à travers son casque VR ou AR. Ces enregistrements fournissent une perspective subjective de l'expérience.
- **Graphes de scène (Scene Graphs)** : Représentation structurée des objets, acteurs et éléments d'environnement dans le monde virtuel sous forme de graphes. Elle permet une analyse fine des relations et interactions dans l'espace virtuel.
- **Données comportementales** : Données capturant les interactions de l'utilisateur (gestes, choix, dialogues, déplacements). Elles complètent l'expérience perçue et enrichissent les requêtes de recherche contextuelle.

L'intégration des MVR dans les systèmes de *Multimedia Information Retrieval* permettrait à un utilisateur de retrouver des instants précis d'une session passée dans un monde virtuel, en s'appuyant non seulement sur les aspects visuels, mais aussi sur les contextes d'interaction, les séquences d'action ou les intentions exprimées.

Dans cette optique, les auteurs de [36] proposent également une architecture dédiée au traitement des enregistrements issus du métavers, intitulée **Process Framework for Metaverse Recordings (PFMR)** (voir figure 14) . Ce cadre conceptuel étend les principes d'un pipeline de traitement vidéo générique pour l'adapter aux spécificités des MVR. Il se compose de plusieurs étapes interconnectées, modélisées sous forme de graphe orienté :

$$PFMR = (N, E)$$

où les nœuds N correspondent aux différentes étapes du processus :

$$N \in \{SA, FE, DM, FF, I\}$$

et les arêtes E définissent les relations entre ces étapes

- **Structure Analysis (SA)** : Première étape, elle segmente les contenus MCO (Metaverse Captured Objects) en unités significatives, optimisant ainsi leur pertinence lors des phases ultérieures. L'ensemble des segments constitue le MVR :

$$MVR = \sum Segments$$

- **Feature Extraction (FE)** : Cette phase applique plusieurs méthodes d'extraction de caractéristiques pour produire des descripteurs de bas niveau issus des formats d'entrée. Ces caractéristiques de base sont représentées par :

$$F = F_{MCO} \wedge F_{SRD} \wedge F_{PD}$$

Elles nécessitent un affinage par les étapes ultérieures.

- **Data Mining (DM)** : À partir des caractéristiques extraites, cette étape vise à détecter des motifs sémantiques ou comportementaux comme des interactions typiques, des événements ou des structures répétitives. Par exemple, la co-détection d'un poisson et d'une canne

à pêche peut être interprétée comme l'activité de pêche. Cette transformation est formalisée par :

$$Func(FVA, FVB) = FVC$$

- **Feature Fusion (FF)** : Elle combine les résultats issus de différentes analyses ou classifications en un vecteur unique, souvent selon des règles définies (seuils de confiance, pondérations, etc.), ce qui améliore la qualité des descripteurs retenus :

$$Func(FVA, FVB) = FVAB$$

- **Indexation (I)** : Dernière étape, elle stocke les représentations obtenues dans une base d'indexation, afin de les rendre accessibles dans des systèmes MIR.

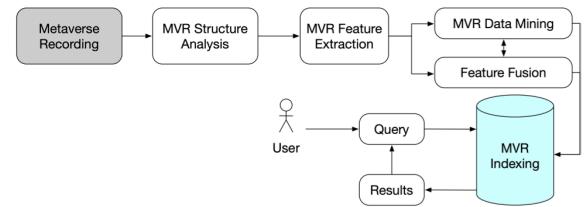


FIGURE 14. PFMR présenté en [36]

Ce modèle permet ainsi de structurer efficacement le traitement de contenus issus du métavers, en assurant une extraction d'information multimodale, orientée sémantique et prête pour l'indexation.

B. Réalité augmentée (AR)

En plus des métavers, la **réalité augmentée (AR)** émerge comme une autre forme de multimédia innovante. Contrairement à la réalité virtuelle (VR), qui immerge l'utilisateur dans un monde entièrement virtuel, l'AR superpose des éléments numériques à l'environnement réel. L'AR peut utiliser des dispositifs comme les smartphones, les lunettes AR ou même des interfaces plus avancées comme les lentilles connectées.

Les applications de l'AR incluent la visualisation de données en temps réel, la navigation assistée, ou l'intégration d'éléments interactifs dans des environnements physiques. Les systèmes de recherche multimédia pour l'AR se basent sur des *représentations d'objets 3D*, des *annotations spatiales*, ou encore des *données contextuelles* (localisation, mouvement, activité) pour améliorer l'expérience d'interaction.

Par exemple, Shakeri et al. [63] introduisent un moteur de recherche sémantique spécifiquement conçu pour les domaines de la réalité augmentée (AR), basé sur une **ontologie dédiée à l'AR**. Le but de leur approche est de surmonter les limites des moteurs de recherche classiques, qui peinent à produire des résultats pertinents lorsque les utilisateurs formulent des requêtes variées ou approximatives. En AR, où les concepts sont nombreux et parfois ambigus (ex. : SDKs, dispositifs, cas d'usage), une approche purement textuelle devient insuffisante.

Le moteur proposé repose sur une **méthodologie en trois étapes** :

1) **Prétraitement des données** : Cette étape prépare les documents AR et les requêtes en les nettoyant et en les uniformisant grâce à des techniques de traitement automatique du langage naturel (NLP), telles que la *tokenisation* (découpage en mots) et la suppression des *mots vides* (non informatifs comme "le", "et", etc.). Cela permet de représenter chaque document sous une forme exploitable pour l'étape suivante.

2) **Modélisation sémantique vectorielle (SVSM)** : Chaque document est ensuite représenté sous forme de vecteur dans un espace conceptuel structuré par l'ontologie AR. Cette représentation repose sur :

- les *poids TF-IDF*, qui reflètent l'importance d'un concept donné dans un document ;
- la *similarité sémantique*, calculée comme l'inverse de la distance entre concepts dans le graphe ontologique.

Les concepts les plus généraux (superclasses) de l'ontologie servent de dimensions dans cet espace vectoriel. Ainsi, un document sur la "réalité augmentée en agriculture" sera représenté par des vecteurs pondérés selon des concepts tels que "AR", "agriculture", "SDK", ou encore "tracking".

3) **Clustering sémantique** : Une fois les documents vectorisés, l'algorithme *k-means* est appliqué pour regrouper les documents similaires. Contrairement aux approches classiques qui utilisent la distance euclidienne, Shakeri et al. emploient ici une *distance cosinus*, plus adaptée à la comparaison de contenus textuels. Le résultat est une organisation des documents en clusters thématiques, correspondant à des sous-domaines de l'AR (ex. : AR médicale, AR industrielle, AR mobile, etc.).

Ce système permet une recherche plus pertinente et contextuelle : les résultats sont regroupés selon leur signification sémantique, plutôt que par simple correspondance de mots-clés. Par exemple, une requête vague comme "*AR en agriculture*" pourra renvoyer efficacement les documents décrivant les SDKs utilisés, les dispositifs compatibles ou les cas d'usage associés, même si ces termes n'ont pas été explicitement mentionnés dans la requête.

Cette approche, centrée sur la *compréhension ontologique des contenus*, illustre une direction prometteuse dans l'adaptation des moteurs de recherche multimédia aux contextes riches et dynamiques comme la réalité augmentée.

C. Hologrammes et Contenus 3D

Les **hologrammes** et les **contenus 3D** offrent une nouvelle dimension dans la capture et la représentation des objets et des scènes. Contrairement aux images et vidéos traditionnelles en 2D, les hologrammes permettent une visualisation tridimensionnelle complète, offrant aux utilisateurs la possibilité d'observer les objets sous différents angles et d'interagir avec eux dans un espace virtuel immersif.

La recherche d'**objets 3D** ou de **contenus volumétriques** repose sur des méthodes spécifiques qui analysent des

représentations comme les maillages (*meshes*), les nuages de points (*point clouds*) ou des vues multi-images. Ces contenus nécessitent des algorithmes capables de traiter à la fois la géométrie, la texture, et parfois l'animation ou l'interaction dynamique dans un espace 3D.

Une contribution récente dans ce domaine est le système **3DMSE – An Interactive 3D Media Search Engine** proposé par Lo Brutto et Meli [57]. Ce moteur de recherche 3D propose une interface web interactive permettant la requête, la visualisation et la navigation dans des collections de modèles 3D. Il intègre des stratégies de recherche unimodale, croisée et multimodale, en combinant les représentations de maillages, nuages de points, et d'images multi-vues. Le système repose sur l'approche *MuseHash*, conçue pour l'encodage multimodal des objets 3D, et permet une exploration intuitive de grandes bases de données tridimensionnelles.

Les moteurs comme 3DMSE illustrent l'évolution des systèmes MIR vers la prise en compte de représentations plus complexes et interactives, ouvrant la voie à des cas d'usage avancés en muséographie numérique, simulation médicale ou design industriel.

D. Interfaces cerveau-machine (BCI)

Les **interfaces cerveau-machine (BCI)** représentent un autre type émergent de multimédia qui permet une interaction directe entre le cerveau humain et un ordinateur, sans l'intermédiaire des dispositifs traditionnels (comme un clavier ou une souris). Ces technologies captent les signaux cérébraux (généralement via EEG, MEG ou NIRS) pour permettre de contrôler des appareils numériques à partir de l'activité cérébrale.

Dans le contexte de la recherche d'information multimédia, les **BCI** pourraient théoriquement permettre à un utilisateur de rechercher ou d'interagir avec des systèmes multimédias simplement en pensant à des concepts, idées ou images. Les signaux neurologiques peuvent ainsi être traduits en intentions ou préférences de recherche.

Le travail de McCartney et al. [58] constitue une avancée significative vers la mise en place d'une interface cerveau-ordinateur (BCI) réellement opérationnelle pour la recherche d'images. Les auteurs proposent un cadre d'apprentissage en zero-shot basé sur les signaux EEG, permettant d'identifier l'image exacte qu'un utilisateur est en train de visualiser, sans que cette image n'ait été vue pendant l'entraînement. Contrairement à des approches antérieures centrées sur la reconnaissance de catégories générales, ce système vise une tâche bien plus fine : retrouver une image précise à partir de l'activité cérébrale.

Leur approche repose sur l'extraction conjointe de deux types de caractéristiques issues des images : des descripteurs visuels (issus de réseaux de neurones convolutifs entraînés sur de grandes bases d'images), et des descripteurs sémantiques (provenant de modèles de traitement du langage tels que des embeddings textuels). Ces représentations sont ensuite mises en correspondance avec les signaux EEG au moyen d'un

modèle d'apprentissage supervisé capable de généraliser à de nouveaux stimuli.

Les auteurs valident leur méthode sur deux jeux de données EEG, comprenant chacun plusieurs participants ayant visualisé des centaines d'images photographiques. Ils appliquent une contrainte stricte de zero-shot learning, dans laquelle toutes les images testées sont absentes de la phase d'entraînement. Les résultats montrent que leur système est capable de retrouver avec précision, parmi un ensemble de candidats, l'image visualisée par un utilisateur, uniquement à partir de ses données EEG.

IX. ÉVALUATION EN RECHERCHE D'INFORMATION MULTIMÉDIA (MIR)

L'évaluation en Recherche d'Information Multimédia (MIR) est un aspect fondamental qui permet de mesurer la performance des systèmes de recherche en termes de pertinence des résultats, d'efficacité et d'efficience. Étant donné la diversité des types de données multimédia (images, audio, vidéo, textes, etc.), l'évaluation en MIR est un domaine complexe qui nécessite des méthodes spécifiques adaptées à chaque type de média ainsi qu'aux besoins des utilisateurs. L'objectif de l'évaluation est de garantir que les systèmes de recherche offrent des résultats pertinents, utiles et faciles à comprendre dans des environnements multimodaux.

A. Particularité des systèmes multimodaux

L'évaluation des systèmes multimodaux inclut :

- **Fusion de modalités** : L'efficacité de la fusion de différentes modalités, comme la combinaison de la recherche visuelle et de la recherche textuelle, doit être mesurée pour déterminer la capacité du système à fournir des résultats pertinents à partir de différentes sources d'informations.
- **Évaluation de la complémentarité** : Un système multimodal efficace doit savoir utiliser la complémentarité entre différentes modalités pour améliorer les résultats de la recherche. Par exemple, une recherche multimodale peut combiner la recherche d'images avec une analyse de texte associé pour mieux comprendre le contexte et améliorer la précision.
- **Mesures globales de performance** : Les mesures globales comme la précision et le rappel sont souvent adaptées pour mesurer les performances globales des systèmes multimodaux en tenant compte de la contribution de chaque modalité.

B. Évaluation de la pertinence temporelle et contextuelle

Les données multimédia, en particulier les vidéos, l'audio et le métavers, peuvent être dynamiques et évolutives. L'évaluation doit également tenir compte de la pertinence temporelle et contextuelle. Par exemple, dans la recherche vidéo, la pertinence d'un extrait vidéo peut dépendre non seulement du contenu visuel, mais aussi de son moment dans la séquence temporelle. De même, dans les environnements virtuels ou dans le métavers, la pertinence des informations

peut dépendre du contexte spécifique de l'utilisateur à un instant donné.

C. Évaluation des performances en MIR : vers un cadre formel

Si les métriques traditionnelles de l'IR comme la précision, le rappel ou le nDCG restent utiles, elles doivent être adaptées aux contextes spécifiques de la MIR.

Dans ce cadre, l'article *Performance Evaluation in Multimedia Retrieval* de Sauter et al. [62] introduit une approche systématique et reproductible de l'évaluation en MIR. Les auteurs proposent un **modèle formel** de l'évaluation, composé de six éléments principaux :

- **Collection de test** : corpus multimédia à interroger,
- **Tâches** : objectifs ou besoins d'information à satisfaire,
- **Agents** : entités effectuant la recherche (humains ou automatiques),
- **Exécution de l'évaluation** : session concrète de test,
- **Jugements de pertinence** : évaluation manuelle ou automatique des résultats,
- **Analyse** : calcul et interprétation des métriques de performance.

Pour accompagner ce modèle, les auteurs présentent **DRES** (**Distributed Retrieval Evaluation Server**), une plateforme open-source d'évaluation, conçue pour les évaluations synchrones et asynchrones. Elle a été utilisée avec succès dans des compétitions comme le *Video Browser Showdown (VBS)* et le *Lifelog Search Challenge (LSC)*.

L'approche de Sauter et al. souligne la nécessité de disposer de *cadres d'évaluation reproductibles*, flexibles et adaptés à la multimodalité, tout en permettant une implication humaine plus large dans les processus d'évaluation.

X. CONCLUSION

La recherche d'information multimédia (MIR) a considérablement évolué au fil des deux dernières décennies, passant de méthodes fondées sur des descripteurs manuels et des modèles vectoriels simples à des systèmes intelligents capables d'intégrer des représentations profondes et multimodales. Dans ce travail, nous avons présenté les principales techniques d'indexation et de recherche utilisées pour les médias traditionnels tels que l'image, la vidéo et l'audio, en mettant en évidence l'évolution des approches de la recherche par le contenu aux modèles neuronaux avancés.

En parallèle, l'émergence de nouveaux types de médias – enregistrements du métavers, hologrammes, environnements AR/MR, interfaces cerveau-machine ou encore données IoT – impose des défis inédits en termes de représentation, d'indexation, d'interprétation et d'interaction. Ces formes de contenus nécessitent non seulement des modèles d'analyse spécifiques, mais aussi une redéfinition des critères de pertinence et des méthodes d'évaluation adaptées à leur nature immersive, dynamique et contextuelle.

L'évaluation des systèmes de MIR, quant à elle, s'oriente vers des cadres plus rigoureux, flexibles et reproductibles,

comme l'illustre l'initiative DRES. Cette évolution est indispensable pour garantir la validité des comparaisons entre systèmes et soutenir le développement d'approches adaptées à la diversité croissante des données multimédia.

En somme, la MIR moderne ne peut plus se penser en silos de modalités isolées, mais doit évoluer vers une approche unifiée, ouverte à la multimodalité, à l'interaction humaine, et à la contextualisation intelligente. C'est à cette condition que les systèmes de recherche pourront répondre efficacement aux nouveaux besoins d'information dans des environnements numériques de plus en plus riches et complexes.

RÉFÉRENCES

- [1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, *Learning Transferable Visual Models From Natural Language Supervision*, arXiv preprint arXiv :2103.00020, 2021. Disponible sur <https://arxiv.org/abs/2103.00020>.
- [2] D. H. Ballard et C. M. Brown, *Computer Vision*, Prentice Hall, 1982.
- [3] M. Levine, *Introduction to Neural and Cognitive Modeling*, Lawrence Erlbaum Associates, 1985.
- [4] R. M. Haralick et L. G. Shapiro, *Computer and Robot Vision*, Addison-Wesley, 1993.
- [5] M. D. Flickner et al., *Query by Image and Video Content : The QBIC System*, IEEE Computer, vol. 28, no. 9, pp. 23–32, 1995.
- [6] J. Bach et al., *Virage : An Information Retrieval System for Images*, Proceedings of the 3rd ACM International Conference on Multimedia, pp. 1–8, 1996.
- [7] J. R. Smith et S. F. Chang, *VisualSEEk : A Fully Automated Content-Based Image Querying System*, Proceedings of the ACM International Conference on Multimedia, pp. 87–98, 1997.
- [8] J. Frankel et al., *WebSeer : An Image Search Engine for the World Wide Web*, Proceedings of the 4th International Conference on Multimedia, pp. 1–10, 1996.
- [9] R. Bluijute et al., *Content-Based Image Retrieval in Informix Database*, Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, pp. 1–12, 1999.
- [10] D. Egas et al., *Extending IBM DB2 with Content-Based Retrieval Capabilities*, Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, pp. 13–24, 1999.
- [11] J. Hanjalic et al., *A New Approach to Shot Boundary Detection*, Proceedings of the IEEE International Conference on Image Processing, pp. 1–4, 1997.
- [12] M. Haas et al., *Scene Detection in Video Sequences Using Motion Analysis*, Proceedings of the IEEE International Conference on Image Processing, pp. 1–4, 1997.
- [13] R. Lienhart, *Reliable Transition Detection in Videos*, Proceedings of the IEEE International Conference on Image Processing, vol. 3, pp. 1–4, 2001.
- [14] H. A. Rowley, S. Baluja, et T. Kanade, *Neural Network-Based Face Detection*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 203–208, 1996.
- [15] R. Lew, *ImageScape : A System for Content-Based Image Retrieval*, Proceedings of the IEEE International Conference on Image Processing, vol. 3, pp. 1–4, 2000.
- [16] A. W. M. Smeulders et al., *Content-Based Image Retrieval at the End of the Early Years*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349–1370, 2000.
- [17] J. P. Eakins et al., *Content-Based Image Retrieval : A Report to the JISC Technology Applications Programme*, 2003.
- [18] J. Foote, *Content-Based Retrieval of Music and Audio*, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 6, pp. 1321–1324, 1999.
- [19] D. Forsyth et M. Fleck, *Detecting Faces in Images : A Survey*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 1, pp. 1–8, 1999.
- [20] M. Bosson et al., *Automatic Detection of Pornographic Content in Images*, Proceedings of the IEEE International Conference on Image Processing, vol. 2, pp. 1–4, 2002.
- [21] L. Rowe et A. Jain, *Multimedia Information Retrieval : A Survey*, ACM Computing Surveys, vol. 37, no. 3, pp. 1–35, 2005.
- [22] Spierenburg, S., & Huijsmans, M. (1997). Image clustering for video creation. *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*.
- [23] Sundaram, H., & Bhat, S. (2002). Video skimming using scene classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(8), 617-624.
- [24] Snoek, C.G.M., Worring, M., & Kraaij, W. (2005). Video retrieval using concept-based approaches. *IEEE Transactions on Multimedia*, 7(3), 487–496.
- [25] Chiu, T., & Huang, T. (2005). Texturizing 3D virtual environments using video. *Proceedings of the IEEE International Conference on Multimedia*.
- [26] Uchihashi, T., & Igarashi, T. (1999). Manga-like summarization of video sequences. *Proceedings of the IEEE International Conference on Multimedia*.
- [27] Tian, Q., & Zhang, B. (2002). Optimizing video summarization through keyframe size and position. *Proceedings of the ACM Multimedia Conference*.
- [28] Liu, Y., & Zhang, H. (2004). Efficient image summarization using clustering techniques. *Proceedings of the International Conference on Image Processing*.
- [29] Egas, P., & Lew, M. (1999). High-performance multimedia databases : Similarity search in large image and video databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 547-558.
- [30] Lew, M., & Egas, P. (2000). Efficient indexing of multimedia data : A comparative study. *Journal of Visual Communication and Image Representation*, 11(1), 56-70.
- [31] Scott, D., & Zhang, H. (2003). Entropy-based balancing for k-d trees in high-dimensional multimedia indexing. *IEEE Transactions on Knowledge and Data Engineering*, 15(5), 1234-1245.
- [32] Ye, J., & Xu, W. (2003). Vector quantization for large-scale similarity search in multimedia databases. *Proceedings of the International Conference on Multimedia and Expo*, 222-225.
- [33] Elkwaie, R., & Kabuka, M. (2000). A two-level indexing method for efficient searching in large image databases. *Proceedings of the IEEE International Conference on Image Processing*, 214-217.
- [34] Shao, C., & Zhang, Y. (2003). Efficient nearest neighbor search using invariant features for large multimedia databases. *Proceedings of the IEEE International Conference on Multimedia and Expo*, 763-766.
- [35] Muller, J., & Henrich, D. (2003). Efficient P2P search using data summaries for multimedia content retrieval. *Proceedings of the IEEE International Conference on Peer-to-Peer Computing*, 34-38.
- [36] Patrick Steinert, Stefan Wagenpfeil, Ingo Frommholz, and Matthias L. Hemmje. Integration of Metaverse Recordings in Multimedia Information Retrieval. In *Proceedings of the 2024 13th International Conference on Software and Computer Applications (ICSCA '24)*, pages 137–145. Association for Computing Machinery, New York, NY, USA, 2024. doi:10.1145/3651781.3651802. URL : <https://doi.org/10.1145/3651781.3651802>.
- [37] Universiti Teknologi MARA. *CSC545 : Multimedia Information Retrieval – Course Notes and Materials*. Faculty of Computer and Mathematical Sciences, 2024.
- [38] Lew, Michael and Sebe, Nicu and Djeraba, Chaabane and Jain, Ramesh. "Content-based multimedia information retrieval : State of the art and challenges." *TOMCCAP*, vol. 2, no. 1, pp. 1-19, Feb. 2006. DOI : 10.1145/1126004.1126005.
- [39] Junrong Huan. *Research on the Application of Artificial Intelligence in Image and Text Database Retrieval*. Frontiers in Computing and Intelligent Systems, Vol. 2, No. 1, 2022, pp. 39–44. ISSN : 2832-6024. Hong Kong Metropolitan University, Hong Kong, China.
- [40] S. Bakheet, A. Al-Hamadi, E. Soliman, and M. Heshmat, "Hybrid Bag-of-Visual-Words and FeatureWiz Selection for Content-Based Visual Information Retrieval," *Sensors*, vol. 23, no. 3, p. 1653, 2023. [Online]. Available : <https://doi.org/10.3390/s23031653>

- [41] H. Liu, Z. Wu, M. Yin, *et al.*, "An improved deep hashing model for image retrieval with binary code similarities," *Journal of Big Data*, vol. 11, article 54, 2024. [Online]. Available : <https://doi.org/10.1186/s40537-024-00919-4>
- [42] Weiss, Y., Torralba, A., & Fergus, R. (2008). Spectral Hashing. *Proceedings of the 21st International Conference on Neural Information Processing Systems (NIPS 2008)*.
- [43] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "Deep Image Retrieval : Learning global representations for image search," *arXiv preprint arXiv :1604.01325*, 2016. [Online]. Available : <https://arxiv.org/abs/1604.01325>.
- [44] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-Scale Image Retrieval with Attentive Deep Local Features," *arXiv preprint arXiv :1612.06321*, 2018. [Online]. Available : <https://arxiv.org/abs/1612.06321>.
- [45] J. Sivic and A. Zisserman. Video Google : A text retrieval approach to object matching in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1470–1477, 2003.
- [46] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3) :107–123, 2005.
- [47] C. Snoek, M. Worring, J.-M. Geusebroek, D. Koelma, F. Seinstra, and A. Smeulders, "The Semantic Pathfinder : Using an Authoring Metaphor for Generic Multimedia Indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1678–1689, Nov. 2006. doi:10.1109/TPAMI.2006.212
- [48] Miech, A., Laptev, I., & Sivic, J. (2018). Learning a Text-Video Embedding from Incomplete and Heterogeneous Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2274-2283. doi : 10.1109/CVPR.2018.00242. <http://www.di.ens.fr/willow/research/mee/>
- [49] Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., & Li, T. (2021). CLIP4Clip : An Empirical Study of CLIP for End to End Video Clip Retrieval. *arXiv preprint arXiv :2103.13405*. <https://arxiv.org/abs/2103.13405>
- [50] Ghias, A., Logan, J., Chamberlin, D., & Smith, B. C. (1995). *Query by Humming – Musical Information Retrieval in an Audio Database*. In *Proceedings of the Third ACM International Conference on Multimedia* (pp. 231–236). ACM, San Francisco, California. <https://doi.org/10.1145/217279.215270>
- [51] Yu, Y., Downie, J. S., Chen, L., Oria, V., & Joe, K. (2008). *Searching musical audio datasets by a batch of multi-variant tracks*. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval (MIR '08)* (pp. 121–127). ACM. <https://doi.org/10.1145/1460096.1460117>
- [52] Manocha, P., Badlani, R., Kumar, A., Shah, A., Elizalde, B., & Raj, B. (2018). *Content-based Representations of Audio Using Siamese Neural Networks*. *arXiv preprint arXiv:1710.10974*.
- [53] Haoran Sun, Zimu Wang, Qiuyi Chen, Jianjun Chen, Jia Wang, and Haiyang Zhang. *Language-based Audio Retrieval with Co-Attention Networks*. *arXiv preprint arXiv :2412.20914*, 2024. <https://doi.org/10.48550/arXiv.2412.20914>
- [54] Patrick Steinert, Jan Mischkies, Stefan Wagenpfel, Ingo Frommholtz, and Matthias Hemmje. Information need in metaverse recordings – a field study. *arXiv preprint arXiv :2411.09053*, 2024. doi:10.48550/arXiv.2411.09053. "
- [55] Chen, Xiang and Zhang, Xue and Liu, Hong and Wang, Ming, *Augmented Reality-Based Visual Search with Spatial Context Modeling*, IEEE Transactions on Multimedia, 2023
- [56] Kapoor, Vivek and Sharma, Rohan and Patel, Sanjeev and Kumar, Akash, *AR4Search : A Multimodal Search Platform for Collaborative AR Environments*, Proceedings of the 2024 International Conference on Multimedia, 2024
- [57] Maria Eirini Pegia, Dimitris Georgalis, Nick Pantelidis, Anastasia Mountzidou, Sotiris Diplaris, Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. *3DMSE : An Interactive 3D Media Search Engine*. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24)*, pages 1260–1264. Association for Computing Machinery, New York, NY, USA, 2024. <https://doi.org/10.1145/3652583.3657593>.
- [58] Ben McCartney, Jesús Martínez-del-Rincón, Barry Devereux, and Brian Murphy. Towards a real-world brain-computer interface for image retrieval. *PLOS ONE*, 14(3) :e0214342, 2019. doi:10.1371/journal.pone.0214342.
- [59] Zhang, Tian, Tao, Jianhua, Wang, Yafeng, and Huang, Jun. A Brain-Computer Interface for Video Retrieval Based on User's Mental Intention. *Neurocomputing*, 2023.
- [60] Li, Chen, Zhang, Yuan, and Wang, Xin. SensorMIR : Integrating Environmental Sensor Data into Multimedia Information Retrieval Systems. In *Proceedings of the 2023 International Conference on Internet of Things and Smart Systems (IoTSys)*, pages 102–110, 2023. doi:10.1109/IoTSys.2023.1037235.
- [61] Kim, Jung Hoon, Lee, Minseok, and Park, Jeongmin. MIR-IoT : A Multimedia Information Retrieval System Using Physiological Data and IoT Sensors for Emotion and Stress Detection. *Journal of Ambient Intelligence and Humanized Computing*, 2024. doi:10.1007/s12652-024-03567-3.
- [62] Loris Sauter, Ralph Gasser, Heiko Schuldt, Abraham Bernstein, and Luca Rossetto. *Performance Evaluation in Multimedia Retrieval*. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(1), Article 23, 23 pages, December 2024. Association for
- [63] S. Shakeri, H. Alinejad-Rokny, and R. Ramezani, *AR Search Engine : Semantic Information Retrieval for Augmented Reality Domain, Sustainability*, vol. 14, no. 23, p. 15681, 2022,