Thèse

Présentée à

Télécom Paris

Par

Itheri Yahiaoui

Pour obtenir le diplôme de

Docteur de Télécom Paris

Spécialité : **Image et signal**

Construction automatique de résumés vidéos
Proposition d'une méthode générique d'évaluation

Soutenance 20 Octobre 2003

Devant le jury Composé de :

Mme. Nozha BOUJEMAA	Directeur de recherche	Rapporteur
Mr. Chabane DJERABA	Professeur	Rapporteur
Mr. Francis SCHMITT	Professeur	Examinateur
Mr. Philippe JOLY	Maitre de Conference	Examinateur
Mr. Bernard MERIALDO	Professeur	Examinateur
Mr. Benoit HUET	Maitre de Conference	Examinateur

Abstract

La croissance rapide des documents multimédia, comme par exemple l'énorme flux de vidéos qui se trouvent sur les ordinateurs personnels et autres équipements, nécessite le développement de nombreux outils pour leur manipulation. La création automatique de résumés vidéos est un outil performant qui permet de résumer le contenu général de la vidéo et de ne présenter que les parties les plus pertinentes.

A travers cette thèse, nous proposons une nouvelle approche de construction et d'évaluation automatique des résumés vidéos. Cette approche est basée sur un principe que nous proposons et que nous dénommons «Principe de Reconnaissance Maximale». Ce dernier est dérivé d'une tâche réaliste prédéfinie qui consiste à l'identification de l'origine d'un extrait donné ayant uniquement connaissance d'un résumé. Notre méthode de construction est un processus d'optimisation, par rapport au Principe de Reconnaissance Maximale PRM, qui permet de construire le meilleur résumé possible pour aider l'utilisateur dans l'accomplissement de cette tâche d'identification. Le meilleur résumé est celui qui maximise le nombre de réponses correctes données par l'utilisateur. Cette technique peut être utilisée pour la création de résumés de différents types de média.

Dans cette thèse, nous avons adopté ce PRM pour proposer différentes méthodes de création de résumés selon le ou les média pris en considération. En premier lieu, nous avons présenté une première méthode de construction de résumés vidéos en utilisant uniquement les informations visuelles, puis nous avons étudié différentes autres méthodes de construction multi-vidéos. Ensuite, nous avons illustré l'adaptation de ce principe pour la construction de résumés basés exclusivement sur les informations textuelles. Enfin, nous avons proposé une méthode mixte de construction de résumés vidéo-textuels en combinant conjointement lors du processus d'optimisation les informations visuelles et textuelles.

A mes parents et tous ceux que j'aime

Remerciements

Je tiens tout d'abord à remercier Le Cher Bon Dieu et tous ceux que j'ai pu omettre dans ce qui suit:

Mes premiers remerciements vont au Professeur Bernard Merialdo et Mr. Benoit Huet qui m'ont encadré tout au long de cette thèse.

Je remercie Mr Francis Smith, Professeur à Télécom Paris qui m'a fait l'honneur de présider le jury. Merci également à Mme Nozha Boujemaa, Directrice de recherche à l'Inria Recquoncourt et Mr Chabane Djeraba, Professeur à l'université de Lille qui ont accepté d'être les rappopteurs de cette thèse, et à Mr Philippe Joly, maître de conférence à l'université de Toulouse.

J'adresse mes plus vifs remerciements à Mr. Laidoudi Bouakkar, Emmanuel Garcia et David Mary qui ont relu avec attention et gentillesse certains chapitres de ce rapport.

Merci aussi à mon ami Rochdi Merzouki, pour l'outil de rédaction qui m'a fourni "Scientific Work Place", pour sa gentillesse, et surtout pour ses conseils avisés lors de la période de la rédaction de ce manuscrit.

Un grand merci à mes deux amies Aicha Khenniche et Fatima Zeroual pour leur soutien et leur encouragement quotidien tout au long de ces années de thèse.

Merci à mon amie Salima Bouzoual d'avoir assisté à ma soutenance malgré la longue distance.

Merci à mes chères Wafa Ferdi et Aicha Yahiaoui pour leurs prières.

Enfin, Je remercie mes parents, ma soeur Meriem, mes deux frères Aboubaker et Mohamed Nadjib, ma grande famille et tous mes amis pour leur soutien moral et leur confiance.

Table des matières

Liste des figures

Chapitre 1

Introduction et Motivation

Dans ce chapitre introductif, nous présentons notre travail sur le problème de construction automatique des résumés multimédia. Après la mise en évidence de l'importance et l'utilité des résumés de documents multimédia, nous exposons les différentes contributions de cette thèse: la création d'un résumé optimal par rapport à une tâche prédéfinie, la construction de résumés multi-vidéos, l'application du même principe de construction adapté aux différents médias, et l'utilisation simultanée du texte et de la vidéo pour la réalisation d'un résumé vidéo-textuel. Enfin, nous présentons lagencement des différents chapitres qui constituent la thèse.

1.1 Importance des résumés

Il est intéressant et même très important d'apporter des solutions spécifiques à des problèmes réels, notamment si ces solutions particulières nous permettent de mieux comprendre des domaines plus complexes. Par exemple, trouver un traitement pour la maladie de la pneumonie atypique apparue les mois précédents en Asie et se propageant dans le monde entier, ne permet pas seulement d'épargner des vies humaines, d'offrir la guérison à des centaines de gens et de dissiper la peur répandue parmi les populations mais aussi de comprendre des phénomènes de la nature, de la biologie et d'autres domaines jusqu'à ce jour ignorés.

La construction des résumés vidéos (des documents multimédia en général) est un problème digne de recherche et ceci sans avoir besoin de définir préalablement une application spécifique. Nous désignons par «document multimédia»

un document linéaire (dans le temps) composé de un ou plusieurs médias. Trouver la caverne d'Ali Baba c'est être sûr de se trouver en face d'un grand trésor sans savoir son contenu exact. Mais bénéficier de ce trésor exige la connaissance indispensable de la fameuse formule magique «Sesame ouvre toi». D'une manière analogue, être en possession d'une grande base de données multimédia ne servira pas à grand chose en l'absence de moyens efficaces qui permettent d'accéder à son contenu. Et ceci n'est pas suffisant: des outils d'organisation, de structuration et de recherche d'informations particulières sont primordiaux car une recherche aveugle, non structurée dans de telles bases de données gigantesques sera semblable à la recherche d'une aiguille dans une botte (une meule) de foin. Imaginons la situation où des utilisateurs ne cherchant pas quelque chose de précis mais désirant plutôt obtenir une idée générale ou une vue plus ou moins rapide du contenu global de cette base de données multimédia. Disposer de mécanismes qui résument des quantités colossales des données présentes dans la base de donnée sous une forme réduite, condensée, simplifiée et surtout représentative du contenu essentiel et utile est un grand aboutissement par rapport à l'ensemble des services que ces mécanismes rapportent aux usagers.

Afin de mieux comprendre l'importance de la construction des résumés vidéos, nous présentons l'exemple de la couverture médiatique d'un évènement dans le monde concernant différentes classes de la population. Cet évènement consiste par exemple en une guerre déclenchée entre deux états.

Avec le grand nombre de chaînes satellites diffusant des informations concernant le déroulement des actions militaires, des reportages montrant la situation humaine dégradée, des déclarations, des démentis d'un camp ou de l'autre, le flux de données multimédia concernant ces opérations devient énorme, contenant même plusieurs contradictions. Malheureusement la notion du temps n'est pas extensible en fonction de la quantité des données diffusées parallèlement et l'utilisateur ne peut consacrer toutes ses journées à consulter cette grande multitude d'informations. Les membres du croissant ou de la croix rouge ne sont intéressés que par la situation humanitaire; les stratèges se concentrent plutôt sur le déroulement des actions militaires et les avancements des troupes, les industriels de leur coté se focalisent sur les dégâts accomplis en vue de s'arracher des marchés pour la reconstruction, une fois la guerre terminée, de ce qui a été

détruit, enfin les diplomates retiennent les différents discours, les positions des états ainsi que les résultats des sondages. Il est clair que les besoins diffèrent d'un téléspectateur à l'autre, et chacun est contraint de filtrer lui même l'information qu'il désire au détriment de son temps personnel. A partir de ce simple exemple, nous relevons plusieurs questions concernant la structuration et la classification des informations, le résumé de ces dernières en fonction des besoins spécifiques des utilisateurs ainsi que la durée de leur disponibilité, sans oublier le degré de leur intérêt et leur vision des différents aspects des sujets traités par toutes ces chaînes. En d'autres termes, ce genre de situation impose des mécanismes spécifiques et précis de filtrage des informations et ceci nous interpelle à créer et développer des outils performant pour le traitement des flux de données multimédia d'une manière générale.

Dans cette thèse nous nous concentrons sur la création et le développement d'un outil efficace parmi tous ceux nécessaires pour gérer et manipuler des bases de données multimédia. Cet outil consiste en un mécanisme qui résume le contenu d'un document multimédia. Plus particulièrement nous proposons tout au long de cette thèse des méthodes de construction et d'évaluation automatique de résumés de documents vidéos. J'insiste sur le fait que la réalisation d'un tel outil est d'une grande utilité et importance pour plusieurs raisons: il nous permet de faire une synthèse du contenu d'une vidéo après avoir fait une analyse de ses différents composants, et d'extraire des informations utiles et récapitulatives des sujets traités dans la vidéo ou bien des informations spécifiques à des questions et des propos bien définis. D'autre part la construction de résumés vidéos fait gagner aux utilisateurs un temps considérable et contribue au développement de techniques qui seront utilisées pour des applications nouvelles comme par exemple la télévision interactive.

1.2 Les contributions de cette thèse

Le cœur de cette thèse est cependant de rendre les documents vidéos plus utiles pour des utilisateurs n'ayant pas le temps nécessaire pour les visualiser et avoir une idée globale de leurs contenus respectifs. L'idée ne consiste pas à sélectionner d'une manière arbitraire des segments d'une vidéo originale ou des images clefs

pour les regrouper respectivement sous forme d'un résumé dynamique ou statique sans avoir une méthode objective loin des heuristiques ou des suppositions pour évaluer la qualité de ce résumé. Par contre notre idée de base est de définir une méthode de construction et d'évaluation de résumés vidéos optimaux par rapport à une tâche fixée préalablement. La méthode de sélection des images ou des segments composant le résumé s'inspire de la méthode d'évaluation en essayant d'optimiser la performance du résumé résultant par rapport à la tâche que nous voulons réaliser en stimulant le comportement d'utilisateurs réels lors de l'accomplissement de cette tâche réaliste.

Les contributions essentielles de notre travail peuvent être répertoriées selon quatre axes principaux. Avant de détailler ces éléments, nous exposons brièvement quelques aspects généraux abordés tout au long de ce rapport, en d'autres termes les mots clefs représentant le contenu général de cette thèse. L'objectif principal de notre travail est de proposer une nouvelle méthode de construction et d'évaluation automatique de résumés vidéos. Sachant qu'une vidéo est composée de différents médias Vidéo, Audio et Transcription (sous-titres), nous avons pris en considération lors de cette étude uniquement la vidéo et le texte. Pour la partie vidéo nous avons traité les points suivants:

- La représentation des images composant la séquence par des vecteurs caractéristiques du contenu visuel en utilisant seulement des descriptifs de couleur.
- Le jugement de la similarité visuelle d'une façon entièrement automatique en utilisant des mesures mathématiques.
- La sélection semi-automatique du seuil de similarité le plus adéquat à chaque vidéo considérée en fonction de la représentation des images la composant et la mesure utilisée.
- La construction et l'évaluation de résumés vidéos optimaux par rapport à une tâche que nous définissons et leur présentation aux usagers sous forme statique ou dynamique.
- La construction de résumés multi-vidéos où chaque résumé fait ressortir et met en valeur les spécificités de chaque vidéo par rapport aux autres vidéos traitées.
- L' évaluation de la qualité des résumés résultants par des utilisateurs réels.

En ce qui concerne la partie texte, nous avons abordé ce qui suit:

- L'élimination des mots outils (très communs comme les articles,les prépositions, etc...) et l'élaboration d'une étude probabiliste associant une information contextuelle à chaque mot.
- Pour les transcriptions en langue française, la mise en œuvre d'une étude morphologique pour le calcul des racines de mots (morphèmes) afin que l'étude de similarité ne soit pas stricte (identité).
- La construction automatique de résumés textuels.
- L'évaluation de la qualité des résumés textuels créés toujours par rapport à une tâche que nous définissons.

Enfin, nous avons combiné les deux flux vidéo et texte pour construire et évaluer d'une manière automatique des résumés vidéo-textuels.

1.2.1 Création de résumé optimal pour une tâche prédéfinie

Nous considérons notre processus de construction automatique de résumés vidéos comme étant un problème d'optimisation sous un ensemble de contraintes liées à une tâche prédéfinie. Afin d'optimiser, nous essayons de créer un résumé ayant un contenu maximal par rapport aux besoins dérivés directement par la tâche ainsi que ceux des utilisateurs intéressés par le résumé. Regarder le problème de construction automatique de résumés vidéos sous un angle d'optimisation nous garantit une évaluation objective liée à la tâche prédéfinie.

1.2.2 Construction de résumés multi-vidéos

Notre objectif principal est de construire automatiquement des résumés des documents multimédia. Le plus simple pour le multi-vidéos consiste à construire de façon séparée le résumé de chaque vidéo. Cependant, cette approche ne prend pas en compte les similitudes pouvant apparaître dans les différentes vidéos, et ces résumés pourront donc contenir une certaine redondance. Pour remédier à ce problème, nous proposons une nouvelle approche qui consiste à construire simultanément les résumés des différentes vidéo, en prenant en compte leurs

similitudes, de façon à inclure dans chaque résumé les éléments qui différencient le plus une vidéo par rapport aux autres vidéos traitées.

1.2.3 Réutilisation du principe de construction

Un document multimédia est composé de différents média ayant chacun ses caractéristiques. Généralement, pour la construction d'un résumé multimédia, les médias sont analysés indépendamment puis une politique de combinaison est utilisée pour obtenir un résumé global. A chaque type de média, une méthode de sélection convenable à la nature du média est utilisée pour sélectionner les segments les plus pertinents qui doivent être insérés dans le résumé global. Dans ce travail, nous proposons un Principe de Reconnaissance Maximale qui sera le noyau de chaque méthode de sélection quel que soit le type du média pris en considération. Cette réutilisation du même principe permet une cohérence entre les différents composants du document multimédia et nous permet d'affecter des explications significatives des raisons des choix des éléments composant le résumé final.

1.2.4 Combinaison du texte et de la vidéo

Une autre contribution de cette thèse consiste en l'élaboration d'une méthode de construction de résumé multimédia où les informations visuelles et textuelles sont utilisées conjointement lors de la création du résumé. Ce lien que nous créons entre les deux médias nous permet d'obtenir des résumés composés d'éléments de types différents mais complémentaires. Dans ce cas, l'optimisation du résumé global par rapport à la tâche définie prend en compte les différents média simultanément.

1.3 Structure de la thèse

L'organisation des septs autres chapitres commence par la présentation, dans le deuxième chapitre, d'une étude bibliographique de certains travaux effectués dans le domaine du traitement et de la construction de résumés vidéos, ainsi que d'autres travaux menés dans le domaine de la construction des résumés textuels.

Le Chapitre 3 évoque le principe que nous proposons pour la construction des résumés multimédia ainsi que l'idée intuitive sur laquelle est basé ce principe de reconnaissance maximale. Le quatrième chapitre illustre l'adaptation de notre principe de construction de résumé aux données visuelles d'une seule vidéo à la fois quel que soit son type (documentaire, film, série, journal télévisé, etc...). Le chapitre suivant montre nos méthodologies de construction de résumés multi-vidéos ainsi qu'une comparaison de nos méthodes avec d'autres méthodes existantes dans d'autres travaux de recherche dans le domaine de la création des résumés. Dans le même chapitre nous comparons les résultats de notre méthode d'évaluation automatique avec des évaluations réalisées par des utilisateurs réels. Le sixième chapitre de cette thèse expose l'application de notre principe de reconnaissance maximale aux données textuelles associées aux vidéos traitées. L'avant-dernier chapitre représente une autre contribution de ce travail qui consiste à combiner le texte et la vidéo pour construire un résumé global des flux vidéo et textuel. Enfin nous concluons par un bilan général de ce travail, une analyse des résultats ainsi que quelques perspectives.

Cette thèse peut être lue de plusieurs façons: une fois que le chapitre trois est lu pour comprendre notre principe de base de sélection et de construction de résumés, si vous vous intéressez uniquement à la partie vidéo, vous pouvez lire immédiatement le chapitre 4 et le 5 si vous vous intéressez aussi aux multi-vidéos. Par contre si vous vous intéressez uniquement au texte, vous pouvez lire directement le chapitre 6. Afin de mieux comprendre la combinaison de la vidéo et le texte et la représentation finale de nos résumés multimédia, je vous conseille de lire le chapitre 7. Enfin, pour avoir une idée globale des travaux effectués dans le domaine de la construction des résumés ainsi que les raisons et les motivations de nos propositions, lisez le premier et le deuxième chapitre.

Chapitre 2

Etat de l'Art des Processus de Construction de Résumés

Dans ce chapitre, nous définissons les concepts utiles pour notre travail de recherche comme par exemple la définition des résumés, des différents types, leur nécessité et utilité. Ensuite nous présentons les différents travaux de recherche effectués dans le domaine de la construction de résumés vidéos ainsi que quelques travaux concernant les résumés textuels.

2.1 Introduction

Avec le développement des technologies récentes, de nombreuses familles possèdent de nos jours un caméscope, et ont pris l'habitude de filmer certains événements (fêtes, etc...). D'autres moyens de création, ou de sauvegarde de documents vidéos existent ou sont en plein essor, comme les DVD, Internet, etc... Nous pouvons citer aussi les documents vidéos créés à partir d'enregistrements satellites, les enregistrements des appareils médicaux ou des caméras de surveillance, ainsi que les flux multimédia diffusés par des milliers de chaînes télévisées, ou présents sur le Net. De plus, la technologie numérique permet la construction de vidéos en utilisant des images statiques, et vice versa. Paradoxalement, la sauvegarde de documents audio et vidéo augmente, mais il devient de plus en plus difficile de trouver le temps pour revoir ces documents. Ces quantités énormes de données multimédia ont de loin dépassé la capacité que nous avons à toutes les traiter, et en tirer avantage. Cette situation s'accentue au fil du temps, et de nouveaux

outils plus «intelligents» pour faire face à ce problème deviennent indispensables.

La création automatique de résumés vidéo est un outil performant qui permet de résumer le contenu général de la vidéo et de présenter les parties les plus pertinentes sous forme d'une séquence audiovisuelle ou d'un ensemble d'images représentatives. Les résumés vidéos permettent d'avoir rapidement une idée sur le contenu de très grandes bases de vidéos, sans nécessiter la visualisation et l'interprétation de l'ensemble des vidéos. Cela permet aussi de juger et d'évaluer la pertinence d'un document multimédia par rapport aux autres.

2.2 Définition

Le résumé d'un document multimédia est une version réduite du document original. Cette version est censée comporter les éléments les plus pertinents et les plus représentatifs du contenu du document original. La qualité du résumé peut être mesurée selon différents facteurs. Nous regroupons ces divers facteurs suivant trois axes principaux [MM99].

L'intention: Ce facteur décrit le potentiel d'utilisation du résumé. Ce dernier peut être indicatif, instructif ou appréciatif. Les résumés indicatifs sont construits tels que, soit ils procurent juste les informations nécessaires pour juger la pertinence des documents originaux, soit ils donnent une brève indication des principaux sujets abordés dans les documents. D'autre part, les résumés instructifs sont des résumés qui peuvent servir de substituts aux documents originaux. Ces résumés doivent retenir des détails importants en réduisant le taux d'informations présentées à l'utilisateur. Par ailleurs, les résumés appréciatifs sont ceux qui capturent le point de vue du concepteur du document original par rapport à un sujet ou un thème donné.

La portée: Selon sa portée par rapport au document original, le résumé peut être générique ou spécifique (selon la demande). Si un document original porte sur un ou plusieurs thèmes, alors le résumé générique est celui pour lequel le processus de construction prend en considération le thème ou les thèmes principaux du document original sans aucune exception. Cependant le résumé spécifique (personnel) est construit aux alentours d'un sujet d'intérêt particulier indiqué par la personne à qui est destiné ce résumé.

La couverture: Un résumé peut être construit à partir d'un seul document comme il peut être construit en se basant sur plusieurs documents présentant un sujet commun.

2.3 Résumés vidéos

Différents travaux de recherche ont traité le problème de construction des résumés vidéos en apportant diverses solutions et propositions, malgré que ce soit un domaine de recherche assez récent mais en plein essor. Par ailleurs, la construction de résumés textes a été amplement étudiée dans le domaine du traitement du langage naturel.

Il est évident que la génération des résumés idéaux nécessite une compréhension approfondie du texte. Ceci est actuellement réalisable uniquement pour des documents de domaines particuliers (comme les rapports médicaux). Pour un texte traitant un sujet d'un domaine quelconque, les systèmes peuvent produire des résumés textes de qualité raisonnable en utilisant des méthodes d'analyse du langage naturel combinées à des critères statistiques.

Les approches proposées dans ce domaine seront passées en revue dans la section 2.5. La stratégie principale consiste en une segmentation du texte en un ensemble d'unités élémentaires, puis en une sélection des éléments qui apparaissent représenter plus d'information. Cette stratégie est la base de plusieurs approches de construction de résumés multimédia.

La classification des techniques de construction de résumés vidéos peut être faite selon différents axes: les média utilisés (Vidéo, Audio, Texte), le type du document multimédia (structuré ou non), l'approche de création utilisée (statistique, selon certains critères), le nombre de documents considérés (mono ou multi-vidéos), et la prise en compte des besoins des utilisateurs (intervention éventuelle de ces derniers).

2.3.1 Utilisation exclusive des informations visuelles

La plupart des systèmes cités dans la littérature [HJ99] [GB99] [UF99] [DDAK99] utilisent uniquement le flux vidéo en faisant l'extraction des images [CN00] [CSL99]

ou des segments représentatifs. Les images représentatives (images-clés) d'une vidéo donnée peuvent être sélectionnées à des intervalles de temps uniformes, ou bien en fonction du contenu de la vidéo.

Diverses méthodes ont été proposées. Ces dernières peuvent être groupées en différentes catégories: les méthodes basées sur un découpage en plans, les méthodes basées sur une classification, les méthodes basées sur un sous-échantillonnage, ainsi que des méthodes basées sur d'autres alternatives.

2.3.1.1 Approches basées sur un découpage en plans

Beaucoup de travaux [CI02] se sont concentrés sur le découpage de la vidéo en un ensemble de plans, et la recherche subséquente d'un nombre d'images représentatives du contenu de chaque plan détecté. Les algorithmes de détection de plans exploitent le manque de continuité qui accompagne le changement de la prise de vue de la caméra et de l'édition. Quelques exemples de méthodologies de détection de plans seront présentés dans la section 3.4.2.

Une fois le découpage de la vidéo en plans effectué selon l'une ou l'autre méthode existante, plusieurs suggestions ont été faites pour la sélection d'une image caractéristique du plan: la première image [TAOS93], la dernière image, l'image médiane, l'image la plus proche de l'image moyenne du plan, la première image nette du plan [GFT97], les deux images les plus différentes, etc...Lienhart et al. [LPE99] et Ueda et al. [UMY91] ont représenté chaque plan par ses première et dernière images. Taniguchi [TAT97] a généré une image composite pour représenter le plan avec les mouvements de la caméra. Ferman et al. [FT97] ont établis une classification des images de chaque plan. L'image la plus proche du centre de la plus grande classe a été sélectionnée comme étant l'image représentative du plan.

Le nombre d'images représentatives extraites d'un plan donné peut être prédéfini, comme il peut dépendre de la longueur du plan ainsi que de son contenu visuel et dynamique [ZLSW95]. D'autres travaux proposent de sélectionner un certain nombre de plans selon différents critères avant de choisir les images qui représenteront leur contenu dans le résumé construit, par exemple:

- Uchihachi et Foote [UF99] ont proposé une méthode de construction de résumés vidéos sous forme d'une figure compacte composée de plusieurs imagettes de différentes tailles. Leur méthode de sélection d'images représentatives de la vidéo est basée sur le calcul de l'importance de chaque plan, ensuite l'élimination des plans inintéressants et très redondants. Ce sont donc seulement les images représentatives des plans les plus importants qui sont insérées dans le résumé. Les dimensions de ces images sont modifiées en fonction de leurs importances respectives.
- Dufaux [Duf00] a décrit une technique d'extraction automatique d'une seule image-clé représentative d'une séquence vidéo destinée à un système de recherche de vidéo sur le Web. Pour chaque vidéo retournée par la requête, une imagette représentative du contenu de cette vidéo est affichée. La technique proposée est divisée en trois étapes: détection des plans, sélection du plan le plus pertinent, enfin sélection d'une image représentative appartenant au plan choisi. La sélection du plan et de l'image-clé est basée sur la mesure du mouvement, l'activité spatiale et la présence de personnes. Cette dernière est déterminée par la détection de la couleur de la peau ainsi que la détection des visages.

2.3.1.2 Approches basées sur une classification

L'emploi d'une phase de découpage en plans permet de diminuer la quantité d'information considérée dans la phase de sélection des images représentatives de la vidéo. Cependant, le choix devient local, c'est-à-dire à l'intérieur d'un seul plan sans considérer le reste de la vidéo. Si différents plans, qui se retrouvent à des emplacements espacés dans la vidéo, ont des contenus très similaires alors les images sélectionnées de ces derniers seront certainement assez semblables. Ce qui provoquera une redondance d'information dans le résumé de la vidéo contenant ces images représentatives.

Dans le but de choisir des images représentatives qui sont différentes les unes des autres et qui représentent bien le contenu de la vidéo, il est intéressant de comparer toutes les images de la vidéo entre elles. Quelques chercheurs ont proposé de faire une classification globale de l'ensemble des images, puis de sélectionner une image par classe afin d'être insérée dans le résumé comme étant une image

représentative du contenu visuel de cette classe.

Le nombre de classes construites dépend du nombre d'images qu'on veut insérer dans le résumé. Dans quelques travaux, ce sont les utilisateurs qui ont procuré le nombre d'images représentatives désirées (composant le résumé final). Dans d'autres, ils ont fourni quelques seuils (par exemple la distance de similarité entre les différentes images ou une approximation des erreurs) utilisés lors du processus de la génération des images représentatives.

Parmi les méthodes proposées suivant cette approche, nous mentionnons:

- Girgensohn et Boreczky [GB99] ont proposé une nouvelle technique de sélection d'images représentatives d'une séquence vidéo donnée. Ces images clefs sont utilisées pour distinguer une vidéo parmi d'autres, pour construire des résumés visuels ainsi que pour fournir des points d'accès aux vidéos. Leur méthode consiste à effectuer une classification hiérarchique de l'ensemble des images, ensuite de sélectionner une image représentative de chaque classe en utilisant des contraintes temporelles après avoirs éliminé les classes jugées pas très importantes. La technique a été ensuite combinée avec une interface de navigation pour fournir un meilleur moyen d'accès à une base de donnée comportant plusieurs vidéos.
- Gong et al. [GL00b] [GL00a] ont proposé une technique de construction de résumés vidéos basée sur la décomposition en valeurs singulières. Chaque image de la séquence vidéo a été représentée par un vecteur caractéristique. Elle a été divisée en neuf régions où chaque région est représentée par un histogramme de couleur tri-dimensionnel calculé dans l'espace RGB. Ces neuf histogrammes ont été concaténés pour former le vecteur caractéristique. Ils ont créé une matrice composée des vecteurs caractéristiques sur laquelle ils ont appliqué la décomposition en valeurs singulières. Cette SVD leur a permis de dériver un espace de caractéristiques plus affiné afin de mieux classifier les images selon leur similarité visuelle. Elle a permis aussi de définir une mesure pour calculer le taux du contenu visuel de chaque classe d'images en utilisant le degré de changement visuel. Dans le nouvel espace affiné, ils ont identifié la classe d'images la plus statique pour la considérer comme l'unité du contenu et ils ont utilisé la valeur du contenu calculée à partir de cette classe comme étant un seuil pour classifier le reste des

images. Une fois la classification terminée, ils ont composé le résumé en fonction de la taille désirée par l'utilisateur. Leur approche assure que le résumé résultant contient peu de redondances et accorde la même importance aux classes qui ont le même taux de contenu visuel.

2.3.1.3 Combinaison du découpage en plans avec une classification

Chacune des approches présentées jusque là a ses avantages et ses inconvénients. Les approches basées sur un découpage en plans sont moins coûteuses en temps de calcul que celles basées sur une classification globale, cependant elles peuvent causer des redondances au sein du résumé ou paradoxalement une perte d'information. Afin de trouver un bon compromis entre ces deux types d'approches, quelques chercheurs [HT97] [DDAK99] [ZZC96] [YBL96] ont proposé de les combiner dans le but de bénéficier de leurs avantages.

Parmi les travaux réalisés dans ce sens nous citons:

- Zhong et al. [ZZC96] ont proposé une technique de classification hiérarchique pour la vidéo. Les plans représentatifs de chaque niveau de la hiérarchie forment des pointeurs sur ce niveau. Ces pointeurs permettent la navigation à travers la vidéo. L'entrée de ce système consiste en une vidéo formée d'une succession de scènes (par exemple les journaux télévisés). Ils ont commencé par une détection de plans pour définir les différents plans constituant la vidéo. Une image représentative de chaque plan a été choisie. Ensuite les vecteurs caractéristiques de ces images ont été utilisés dans le processus de classification hiérarchique. Ils ont associé à chaque image les vecteurs caractéristiques suivants: l'histogramme de couleur, l'histogramme d'orientation des vecteurs de mouvement, la moyenne et la variance du mouvement. Ils ont utilisé un algorithme de classification de type K-means flou. Par ailleurs, l'utilisation d'une seule image par plan ne permet pas de capturer l'information temporelle de l'événement.
- Zhuang et al. [ZRHM98] ont proposé une autre technique de sélection d'images-clés représentatives d'une vidéo en utilisant une classification non supervisée. Après une division en plans, les images de chacun des plans ont été classifiées selon leur contenu visuel en plusieurs classes en fonction

d'un seuil de similarité prédéfini. Cette classification se fait de la manière suivante: la première image est incluse dans une première classe, ensuite des valeurs de similarité sont calculées entre une future image et les centroïdes des classes déjà existantes. Si la valeur maximale est plus petite que le seuil, une nouvelle classe est créée en incluant cette image sinon l'image est ajoutée à la classe qui contient des images qui lui sont visuellement similaires, et ainsi de suite jusqu'à la dernière image du plan. Une fois la classification terminée, ils sélectionnent une image de chaque classe sauf celles de taille réduite. Les images sélectionnées sont les plus proches des centroïdes.

2.3.1.4 Approches basées sur un sous-échantillonnage

En général la sélection des images représentatives est faite d'une telle manière que les images les plus caractéristiques de la vidéo sont extraites. Néanmoins, la sélection d'un ensemble de bonnes images représentatives de la vidéo originale est subjective et représente un vrai défi.

La création d'un résumé vidéo sous forme d'une collection d'images représentatives du contenu visuel de la vidéo permet à l'utilisateur d'avoir une idée globale de la vidéo entière en utilisant un petit nombre d'images. Cependant, un ensemble d'images statiques ne permet pas de capturer la dynamique et la continuité des images d'une séquence vidéo.

Dans le but de construire un résumé vidéo dynamique qui soit sous forme d'une séquence visuelle réduite extraite de la vidéo originale au lieu d'un ensemble statique d'images-clés (représentatives), différentes méthodologies ont été proposées [ML00] [LGA+99]. Ces méthodologies procèdent, soit en regroupant des segments de la vidéo relatifs à des images clefs présélectionnées par l'une des méthodes d'extraction d'images-clés, soit en faisant un sous-échantillonnage.

Comme exemple de ce deuxième type de méthodes, nous trouvons:

- Nam et Tawfik [NT99] ont proposé une procédure de construction de résumés vidéos qui produit un résumé dynamique de la vidéo originale. Leur approche dépend d'un sous-échantillonnage non-linéaire adaptatif à la vidéo. Le taux local de sous-échantillonnage est directement proportionnel à l'intensité

d'activité visuelle dans les sous-plans localisés dans la vidéo. Pour présenter le résumé aux utilisateurs, une interpolation linéaire a été utilisée pour augmenter le nombre d'images dans le but de préserver une longueur et un taux d'activité relative à chaque sous-plan.

- DeMenthon et al. [DKD98] ont proposé de représenter la vidéo sous forme d'une trajectoire dans un espace de caractéristiques de grande dimension. Cette courbe a été analysée par des méthodes similaires à celles développées pour les courbes planes. Ils ont découpé la courbe avec un algorithme de découpage récursif qui vérifie à chaque étape la dimension du segment à découper. Les résultats de cet algorithme ont été structurés sous forme hiérarchique où chaque niveau correspond à un taux de sous-échantillonnage. Les images de la vidéo qui correspondent aux points de coupure sur la courbe d'un niveau donné de la hiérarchie sont considérées comme des images clefs. Ces images sont affichées dans un ordre séquentiel à une fréquence de 30 images par seconde.

2.3.1.5 Diverses alternatives

Il existe d'autres méthodes [VL98b] de sélection d'images clefs n'appartenant à aucune des catégories présentées antérieurement, par exemples:

- Chiu et al. [CGP+00] ont proposé un algorithme génétique pour la segmentation des vidéos basée sur une fonction de similarité d'images contiguës. Un chromosome est une chaîne de 0 et de 1 où chaque 0 correspond à une image de la vidéo et les 1 correspondent aux images considérées comme des limites des segments. Le nombre de 1 (nombre de coupures) est un paramètre variable dans le processus de segmentation. La fonction de sélection est basée sur le calcul de la similarité des images. Cette dernière est calculée à base des différences des histogrammes de couleurs. Ils se sont inspirés ensuite des travaux de Foote et al. [UF99] pour définir l'importance de chaque segment. Enfin ils ont construit un résumé composé des images qui délimitent les plans plus la première image de la vidéo.
- Stefanidis et al. [SPD00] ont proposé une approche de construction d'un résumé vidéo basée sur l'analyse des trajectoires des objets appartenant

à cette vidéo. Leur travail consiste à identifier lors des trajectoires des objets des nœuds comme étant des points critiques permettant de décrire le comportement d'un objet à travers un segment de la vidéo. Les instants correspondant à ces nœuds sont utilisés pour sélectionner les images qui constituent un résumé vidéo décrivant le comportement d'un objet dans un segment vidéo. L'analyse des positions relatives aux objets intéressants dans la vidéo permet la sélection d'autres images représentatives.

- Pope et al. [PKSW98] ont décrit un autre schéma de construction de résumés vidéos en utilisant les mosaïques et les trajectoires de mouvement des objets. La mosaïque est une image construite par l'assemblage des images composant un plan de la vidéo. Les objets en mouvement ont été ensuite placés dans la mosaïque pour créer une représentation alternative de la vidéo. Ce schéma a été utilisé dans une application de recherche de clips dans une base de données et de visualisation de scènes vidéos. Chaque vidéo a été segmentée en plusieurs clips après une détection des coupures des scènes. Le contenu statique de chaque clip a été résumé par une mosaïque, et le contenu dynamique par la segmentation et le suivi du mouvement des objets. Le contenu dynamique a recouvert la mosaïque pour fournir un mécanisme de détection d'évènements sémantiques (comme par exemple les collisions, poursuites, les activités particulières). Le schéma présenté consiste en une représentation hiérarchique où la racine comporte la vidéo entière et les feuilles renferment les clips auxquels plusieurs informations sont associées.

2.3.2 Construction de résumés par combinaison de différents média

Une séquence vidéo n'est pas seulement une collection d'images, c'est aussi une évolution de liens spatio-temporels où une image constitue au mieux un maillon de cette chaîne. Afin de capturer ceci, nous avons besoin de représentations significatives, compactes et agiles de cette évolution. Différentes caractéristiques du flux vidéo doivent être utilisées simultanément; la vidéo elle-même ainsi que le signal audio (parole, musique, bruit, etc...) et de l'information textuelle contenue

dans le télé texte (sous-titres).

Différents projets de recherche [LPE97] [IL98] construisent un résumé d'images en mouvement en se basant conjointement sur le flux vidéo ainsi que le flux audio, ensuite ces informations sont enrichies par l'extraction d'évènements spéciaux additionnels.

Ces systèmes sélectionnent des extraits en fonction de leur contenu en utilisant beaucoup de techniques d'analyse du domaine de la vision par ordinateur, l'analyse de la parole et le traitement du langage naturel pour construire un résumé sous forme d'une séquence audiovisuelle.

Nous avons sélectionné les exemples suivants:

- Kanade et al. [SK97a] ont travaillé sur le sujet d'une librairie de vidéo intelligente où l'utilisateur est doté d'une interface de recherche dans une base de données. Ils ont travaillé en particulier sur la construction de résumés audiovisuels. En premier, ils ont segmenté la vidéo en utilisant des histogrammes de couleur. Les segments vidéos résultants ont été ensuite associés à la transcription du flux audio générée par le système de reconnaissance de parole Sphinx-II. La technique du TF.IDF (terme de fréquence, fréquence du document inverse) a été appliquée à la transcription de l'audio pour retrouver les mots et les phrases les plus pertinentes. En plus, l'importance relative de chaque segment vidéo a été calculée en utilisant l'analyse du mouvement de la caméra et la reconnaissance des visages. Finalement, ils ont combiné les phrases sélectionnées (l'audio) et les segments vidéos pour produire un résumé de la vidéo originale sous forme d'une séquence audiovisuelle.
- Maybury et al. [MM97] ont examiné différentes manières d'implémenter un processus de sélection afin de former un résumé. Ils ont présenté des techniques de calcul simples, des techniques plus élaborées basées sur l'extraction de caractéristiques à partir de l'audio et de la vidéo, et finalement des techniques structurelles qui prennent en compte les mouvements de la caméra et la structure du contenu comme par exemple la structure des informations difusées sur la télévision. Leur implémentation réelle utilise des caractéristiques statistiques de la transcription du flux audio pour sélectionner le contenu, dont les mots ont les plus grandes fréquences.

2.3.3 Construction de résumés basée sur la structure des documents

Plusieurs chercheurs [KTIA98] [Mer97] ont exploité la structure de vidéos particulières pour proposer des méthodes de construction de résumés spécifiques à ce genre de documents. Comme exemple de vidéos, ils ont utilisé les émissions sportives, les journaux télévisés, les présentations vidéos, etc...

Pour illustrer ces méthodes nous avons sélectionné les articles suivants:

- Huang et al. [HLR+99] ont décrit une méthode de construction automatique d'un résumé compact et significatif de données multimédia qui peut être utilisé comme une table d'indexes efficace pour une recherche non linéaire dans une grande base de données. Leur méthode a été appliquée aux journaux télévisés en utilisant les trois média, la vidéo, l'audio et le texte. Chaque journal a été représenté par une représentation hiérarchique, la racine comporte le journal complet, le premier niveau sépare les infos des annonces publicitaires en combinant des informations audio et vidéos, le niveau d'après divise les infos en parties présentées par le présentateur principal et celles présentées par les reporters en utilisant des techniques de reconnaissance de parole. Ensuite, en utilisant des informations sémantiques déduites d'une analyse des informations audio et texte, d'autres segmentations ont été mises en œuvre pour séparer l'introduction du reporter et le reportage lui même.
- Yueng et al. [YYWL95] ont proposé un Graphe de Transition de Scènes (GTS) comme modèle d'organisation et de navigation dans la vidéo. Le GTS est une représentation dans laquelle les plans composant la vidéo sont organisés en un ensemble de groupes et les transitions entre les plans sont des indications du contenu de la vidéo. Par exemple, faisons l'hypothèse qu'une séquence vidéo est constituée d'une scène de dialogue suivie d'une scène d'action puis une autre scène de dialogue. Chacune de ces scènes est composée de plusieurs plans. Une scène de dialogue aura typiquement deux ou trois prises de vue de la caméra, chaque prise de vue représente un plan. Un simple découpage de la vidéo en un ensemble de plans ne permettra pas le repérage des dialogues. Cependant, si les images représentatives de

chaque plan sont examinées, la scène de dialogue peut être classifiée en deux/trois groupes de plans (un groupe pour chaque prise de vue). Les auteurs ont utilisé cette intuition pour classifier les plans et créer un graphe de transition de scènes où chaque nœud représente un groupe de plans. L'observation de ce graphe montre clairement que l'événement contenu dans une scène est un dialogue ou une action. Il fournit un bon résumé du contenu de la vidéo. Ceci fonctionne bien pour un contenu structuré. Pour les vidéos non structurées, les hypothèses sur lesquelles repose le GTS sont invalides et il ne peut être appliqué. D'une façon intéressante, les auteurs ont discuté l'utilisation d'une seule image représentative par plan [YL95b] en montrant que ceci cause une perte de la structure temporelle et par conséquent une mauvaise classification des plans si l'image représentative est mal choisie. Par conséquent, les auteurs ont utilisé plusieurs images représentatives sélectionnées en fonction de la quantité de mouvement dans un plan. Un plan contenant plus d'actions aura plus d'images pour le représenter. C'est un mécanisme qui permet d'avoir un compromis entre la complexité de calcul dans le cas d'utilisation de toutes les images de la vidéo et la perte d'information dans le cas de celle d'une seule image représentative.

- Ju et al. [JBMK98] ont réalisé une détection des segmentations de scènes en utilisant des informations globales du mouvement. Ensuite, ils ont établi une reconnaissance et un suivi automatique des gestes humains simples lors d'une présentation vidéo. Ils ont utilisé ces gestes et ces actions comme annotation de la séquence qui peut être utilisée pour accéder à une version condensée de la représentation vidéo d'une page Web.

2.3.4 Approches statistiques

Les approches basées sur des mesures statistiques ont été souvent utilisées dans le domaine de la construction des résumés textes. Quelques chercheurs [GL01] [FT99] [DDK00] [VL98a] se sont inspirés de ces méthodes pour proposer de nouvelles méthodes de construction de résumés vidéos basées sur des mesures statistiques.

Nous avons choisi les exemples suivants:

- Sahouria et la. [SZ99] ont utilisé l'analyse en composantes principales (PCA) pour réduire la dimension des vecteurs caractéristiques des images de la vidéo afin de faire une description du contenu. Cette description de dimension réduite permet pratiquement l'utilisation de l'ensemble de toutes les images composant la vidéo. Ensuite cette représentation en PCA était utilisée pour deux applications. La première consiste à faire une description de haut niveau des scènes avec sélection d'une image-clé représentative de chaque scène sans détection ni découpage en plans. La deuxième application concerne la classification de séquences vidéo de sport en utilisant les vecteurs de mouvement de chaque image.
- Iyengar et Lippman [IL97b] ont présenté une technique d'analyse et de classification des séquences vidéos en plusieurs catégories. En premier lieu, des programmes de télévision, journaux télévisés et programmes sportifs ont été utilisés. Les vidéos en entrée sont analysées afin d'extraire les informations de mouvement qui sont projetées d'une manière optimale dans un espace d'une seule dimension. Ces données projetées sont ensuite utilisées pour entraîner des Modèles de Markov Cachés. Ces derniers permettent une classification efficace et précise des séquences vidéos traitées. Encore une fois, ils ont utilisé les Modèles de Markov Cachés pour classifier des bandes annonces de plusieurs films afin de détecter si c'est un film d'action ou non. Pour l'entraînement de ces Modèles, ils ont utilisé la durée de chaque plan ainsi que l'énergie du mouvement. Le découpage en plans est effectué en utilisant la distance Kullback-Leibler entre les histogrammes de couleur des images calculés dans l'espace RGB.

2.3.5 Approches basées sur des critères de sélection

Dans le but de construire des résumés vidéos qui répondent aux attentes des utilisateurs sans intervention directe de ces derniers, quelques groupes de recherche [LPsF97] ont essayé d'analyser les aspects qui retiennent le plus l'attention des utilisateurs. Ces aspects ont été souvent formalisés sous forme d'un ensemble de règles. Ces dernières ont été employées dans le processus de génération des

résumés vidéos.

Voici deux parmi les travaux effectués suivant cet axe:

- Lienhart et al. [LPE97] ont travaillé sur les bandes d'annonce de films. Ils ont identifié une liste de propriétés d'une bande annonce de bonne qualité comme par exemple la présence d'objets et de personnes importantes, les évènements d'action, et le dialogue. Leur implémentation a utilisé la détection de la composition de la scène en fonction du mouvement et du contraste. En plus, ils ont détecté le dialogue à partir du flux audio. Ils ont regroupé le contenu sélectionné en préservant l'ordre chronologique des évènements tels qu'ils étaient dans la vidéo originale et en évitant d'inclure les segments pris de la fin de la vidéo originale pour ne pas dévoiler la fin de l'histoire présentée dans le film. Ils ont aussi gardé une synchronisation de la vidéo et de l'audio sans les considérer indépendamment afin de les réutiliser d'une manière flexible dans la bande annonce.
- Sundaram et al. [SC01] travaillent sur la construction de résumés de vidéos à travers une analyse structurelle complètement automatique des flux audio et vidéo. Leurs travaux se concentrent sur trois axes spécifiques: la relation entre la longueur du plan dans un film et son temps de compréhension, l'analyse de la structure syntaxique dans un film et la définition d'une fonction d'utilité d'une séquence.

2.3.6 Prise en compte des besoins des utilisateurs

Afin de créer un bon résumé vidéo, il est très important de garantir que ce dernier comporte peu de redondances, et surtout comprenne des informations pertinentes par rapport aux demandes des utilisateurs. Pour ces raisons, quelques chercheurs [MLZL02] ont décidé de prendre en compte les besoins des utilisateurs, soit comme une entrée dans le processus initial de construction du résumé, soit lors d'une phase de raffinement d'un premier résumé construit selon des méthodes statistiques ou un ensemble de critères prédéfinis par le concepteur du système.

Parmi les travaux impliqués dans ce type d'approches, nous mentionnons:

- Toklu et al. [TLD00] qui ont proposé une approche hybride de construction de résumés vidéos. Leur système produit d'une manière automatique un

résumé multimédia qui est présenté à l'utilisateur à travers une interface lui permettant de modifier le résumé actuel selon ses besoins. La première étape de leur travail consiste à faire une détection de plans, où chaque plan est représenté par une image représentative, la première du plan. Ensuite ils ont augmenté le nombre d'images clefs sélectionnées pour chaque plan. A l'aide du télétexte, ils ont décomposé l'histoire globale de la vidéo traitée en plusieurs sous-histoires élémentaires. Ils divisent le flux audio en différents segments, les paroles, le silence et les bruits. La construction du résumé est basée sur le texte, puis ils ont combiné le reste (image et audio) pour obtenir un résumé multimédia.

- Oh et al. [OH00] qui ont défini une bonne technique de construction de résumés vidéos comme étant celle qui permet de construire différents résumés d'une même vidéo en fonction des besoins des utilisateurs. Pour ceci, ils ont proposé une technique de construction basée sur l'intervention humaine. Ils présentent plusieurs scènes représentatives du contenu global de la vidéo à l'utilisateur. Ce dernier sélectionne parmi ces scènes, celles qui correspondent le plus à ce qu'il demande ou à l'application pour laquelle le résumé résultant est destiné. Une fois que l'utilisateur termine sa sélection, le système retrouve à travers toute la vidéo d'autres scènes qui sont importantes et pertinentes par rapport aux critères présents dans les scènes sélectionnées. Chaque scène était représentée par un couple de valeurs numériques, ce qui facilite la comparaison et la recherche et augmente la performance et l'efficacité.
- Masumitsu et al. [ME00] qui ont proposé une méthode de construction de résumés vidéos en sélectionnant les images ayant un grand score d'importance. Pour assigner des scores à l'ensemble des images la procédure suivante est appliquée. Chaque image est représentée par un vecteur caractéristique, ce vecteur est calculé en se basant sur les couleurs, la texture, les objets et d'autres caractéristiques. Ensuite le vecteur est normalisé en soustrayant la moyenne puis en divisant par la variance. L'ensemble de ces nouveaux vecteurs normalisés représentant les images composant la vidéo sont projetés dans un espace de vecteurs propres en utilisant l'analyse en composantes principales afin de diminuer la corrélation entre les éléments

composant chaque vecteur caractéristique et de réduire la dimension des vecteurs. Un espace additionnel de paramètres potentiels est introduit pour refléter les préférences d'un utilisateur donné. Une image est considérée comme importante pour un utilisateur s'il prend un certain temps à la regarder sinon si cette image est parcourue rapidement, cela implique qu'elle ne l'intéresse pas. Une fois ces deux espaces définis, les vecteurs caractéristiques projetés, les paramètres de préférences et une mesure de similarité des images sont tous combinés pour assigner les scores d'importance à chaque image. Les images de faible score sont éliminées et les autres sont regroupées pour former un résumé correspondant au besoin d'un utilisateur donné.

2.3.7 Construction de résumés multi-vidéos

Alors que la construction d'un résumé d'une vidéo simple reçoit une attention croissante, peu de travaux [YMH01b] [YMH01a] ont au contraire été consacrés au problème de construction de résumés multi-vidéos. Dans le cas du multi-vidéos, il faut prendre en considération différentes contraintes comme la présence d'information redondante dans les différents documents multimédia traités. En fonction des besoins des utilisateurs ainsi que de la tâche à laquelle est destiné le résumé multividéo, ce dernier peut contenir soit les informations appartenant à l'intersection des différentes vidéos comme étant l'information essentielle (par exemple le cas des journaux télévisés), soit le compléments de l'intersection comme étant les informations spécifiques à chaque vidéo (par exemple le cas des séries télévisées).

2.4 Méthodes annexes à la construction automatique des résumés vidéos

La plupart des méthodes [FTM00] [IL97a] de construction des résumés vidéos font appel à des méthodes de pré-traitement. Ces méthodes [Dje02] [YY97] [HYM02] [Dje03] permettent de caractériser les différents média utilisés, de mesurer la similarité des différents éléments composant la vidéo traitée, et de segmenter la

vidéo en plusieurs parties. Dans cette section, nous présentons très brièvement des aspects relatifs aux informations visuelles. Ces aspects sont la caractérisation des images, la mesure de similarité, et la détection des plans.

2.4.1 Mesure de la similarité visuelle des images

Il est important de trouver les caractéristiques appropriées qui peuvent être extraites des images ainsi que les fonctions de similarité dans cet espace de caractéristiques de façon à ce que deux images dont les caractéristiques sont proches sont visuellement similaires par rapport au jugement humain .

Divers descripteurs de l'apparence visuelle sont utilisés pour la recherche par le contenu. Les principaux descripteurs d'images sont la couleur, la texture et la forme. Parmi ces trois descripteurs la couleur est la plus souvent utilisée et donne actuellement les meilleurs résultats. La texture est aussi très utilisée bien qu'elle soit plus complexe à mettre en œuvre que la couleur. La forme est de loin la plus complexe à représenter et à comparer.

La représentation d'images la plus commune du domaine de la recherche d'images dans les bases de données est l'histogramme de couleur. Il est raisonnable que les images qui se ressemblent auront des compositions de couleur similaires et donc il est envisageable que les histogrammes soient intéressants. Les histogrammes de couleur ont d'autres avantages, ils sont invariant par rapport à la taille, à l'échelle et à la rotation des images. De plus ils sont compactes si nous les comparons avec la taille des images. Une fois que les images sont représentés par les histogrammes dans un espace de couleurs, nous pouvons choisir une parmi les différentes mesures de similarité existantes. Swain et Ballard [SB91] introduisent la notion d'intersection d'histogrammes, qui est équivalente à la distance Manhattan (norme L_1) pour des histogrammes normalisés. D'autres mesures souvent utilisées sont les normes Ln et en particulier L_1 et L_2 [JV96], l'estimation χ_2 [DBVF98], la divergence KL [IL98], et la logique floue [Bou00] [LCW03]. Un des problèmes de l'utilisation directe des histogrammes est qu'un petit décalage des valeurs des vecteurs caractéristiques implique une grande différence entre des deux histogrammes en utilisant une des mesures de similarité communes. Les histogrammes de couleurs souffrent d'un manque d'information locale parce que

ces derniers n'enregistrent pas la structure locale des couleurs mais seulement la composition globale. Le vecteur de cohérence de couleurs [PZ96] est une représentation affinée de l'histogramme où les couleurs sont divisées en deux composantes: cohérente et non cohérente. La composante cohérente comporte le nombre de pixels voisins ayant la même couleur que le pixel en cours. La distance entre les histogrammes de ce type dépend des deux composantes. Une autre représentation plus affinée de l'histogramme est le corrélogramme de couleurs [HRKM+97] qui enregistre la probabilité de présence d'une couleur dans un voisinage prédéfinie d'une autre couleur. Une nouvelle représentation des information locales et globales conjointement est nommée les histogrammes de blobs [How98].

Malgré que les histogrammes de couleur sont compacts et efficaces, ils ne sont pas utiles pour tous les types d'images (les images à niveau de gris). Quelques travaux de recherche ont utilisé la texture. Liu et Picard [LP96] ont proposé de décomposer l'image selon trois axes: la périodicité, la direction, et le caractère aléatoire. Ensuite un modèle probabiliste a été construit à travers ces trois axes pour caractériser les images. Ma et Manjunath [MM96] ont utilisé les filtres de Gabor pour la caractérisation des images et la distance Mahanalobis pour le calcul de leur similarité.

D'autres équipes de recherche [LWW00] ont proposé de décrire le contenu des images en utilisant les formes. Smith et Chang [SC97] ont divisé les images en un ensemble de régions ou objets et ont calculé la similarité d'une paire d'images en fonction des formes détectées dans chacune d'elles. Jain et Vailaya [JV96] ont utilisé un histogramme de coins pour représenter les formes dans une image.

Plusieurs systèmes [VXJ98] [VB00] combinent différents descripteurs pour former les vecteurs caractéristiques des images. Cette combinaison permet d'améliorer les performances de la détermination de la similarité visuelle.

Cependant la manière dont la similarité des images est perçue par les humains ne peut être représentée par ces descripteurs de bas niveau. La quantification de la similarité perceptive reste un problème très délicat. Li et al. [LCW03] ont découvert une nouvelle mesure de la similarité visuelle. Ils ont comparé en pratique cette nouvelle mesure avec les mesures de type Minkowski dans une application de recherche d'images ainsi qu'un découpage en plans. Les images ont été représentées selon certaines caractéristiques (couleur, texture, forme) définies par

les auteurs. Leur méthode a enregistré des meilleures performances. L'efficacité de cette méthode peut être expliquée par les théories de similarité dans le domaine de la psychologie cognitive. Ils ont utilisé une très grande quantité de données visuelles afin de retrouver une bonne distance de mesure de similarité visuelle. Après la création d'une grande base de données visuelles de 60000 images de type JPEG à partir des CD et Internet, ils ont appliqué 24 transformations sur chacune des images pour obtenir de nouvelles images qui lui sont similaires. Ils ont utilisé ces nouvelles images (60000*25) pour définir une caractérisation et une mesure de similarité proche de la perception humaine. En utilisant la distance euclidienne, ils ont comparé différentes caractérisations basées sur la couleur, la forme, la texture ou différentes combinaisons de ces descripteurs. La meilleure performance a été enregistrée par la combinaison de la couleur et de la texture (60%). Ensuite, ils ont retrouvé une nouvelle distance de similarité visuelle qui permet d'améliorer les résultats de la distance euclidienne. Pour ceci ils ont fixé la représentation et étudié plusieurs distances en combinant différemment les vecteurs caractéristiques correspondant aux différents descripteurs. Une fois la meilleure représentation d'images pour les calculs de la similarité visuelle en utilisant cette nouvelle mesure définie, un test est effectué sur un ensemble de 1,5 million d'images.

2.4.2 Détection de Plans

La détection de plans dans une séquence vidéo peut être réalisée en utilisant éventuellement le flux compressé (Mpeg) [CSI+02] ou non [BR96] [BGG99].

Parmi les travaux présentés [LMZW02] [JLZ00] [KM02] concernant la détection des plans dans le domaine décompressé, nous citons:

- Shahraray [Sha95] qui a proposé une technique de détection de changement de plans dans laquelle des paires d'images successives à comparer ont été divisées en blocs rectangulaires et des valeurs de mouvements et d'intensité calculées à partir de ces blocs ont été utilisées pour la mise en correspondance. Si deux blocs ont les mêmes valeurs de mouvement et d'intensité, la valeur 0 est renvoyée. Une valeur de 1 indique une grande différence entre les valeurs correspondant aux deux blocs. La similarité cumulée a

été définie comme la somme des similarités (valeurs renvoyées) calculées pour l'ensemble des blocs. Un changement de plan est repéré si la valeur cumulée est plus grande qu'un seuil. Une transition graduelle est détectée lorsque la valeur cumulée est supérieure à un seuil plus petit et que cette valeur avoisine ce petit seuil pour un groupe d'images avant qu'elle tombe en dessous du plateau. L'avantage principal de cette technique est qu'elle combine les paramètres du mouvement et de l'intensité dans le processus d'évaluation du changement de plan. Ce qui fait d'elle une technique assez robuste.

- Zabih et al. [ZMM96] qui ont décrit une technique de détection de transition de plans en utilisant des informations relatives aux coins (bords). L'idée de cette technique découle du fait que durant une coupure ou une transition de type fondu enchaîné, des nouvelles bordures apparaissent loin des positions des anciennes et vice versa. Ils ont utilisé cette observation pour classifier un ensemble de pixels comme étant des coins arrivants ou partants. Les coupures de plans ont été détectées en recherchant les pics locaux dans les différences entre les nouveaux et les anciens pixels de bords. Avant le calcul des pixels de coins, les auteurs ont compensé le mouvement en calculant le mouvement dominant. Cette initiative permet de garantir que les opérations liées à la caméra (par exemple zoom ou panoramique) ne contribuent pas à des fausses détections de plans.

Les algorithmes de détection de plans dans le domaine décompressé sont coûteux en temps de calcul. De plus, les changements de couleur et de mouvement peuvent être extraits directement du flux compressé. La détection de changement de plans dans un flux vidéo compressé comme MPEG est basée sur des statistiques effectuées sur les coefficients DCT ou les paramètres d'estimation de mouvement [YL95a]. Quelques exemples sont le produit scalaire des coefficients DCT des images I comme étant une mesure de similarité des images, le rapport entre le nombre de blocs des images intra codées P et le nombre de blocs des images inter codées, le rapport entre le nombre de blocs des images B prédites en avant et ceux des images prédites en arrière, la variance des coefficients DCT, etc... Quelques approches reconstruisent partiellement les images à partir des coefficients DC et utilisent ensuite des histogrammes représentant ces images pour

détecter les coupures des scènes. L'article de Gargi et al. [GKA98] présente une comparaison de quelques techniques de la détection des changement des plans.

2.5 Synthèse globale

Afin de donner une idée globale et représentative des différentes méthodes de construction de résumés vidéos proposées dans la littérarture, nous avons effectué une analyse des principaux travaux. Pour chaque travail de recherche, nous avons identifier les différents critères sur lesquelles repose la méthodologie proposée. Ensuite nous avons classifié l'ensemble des critères correspondant aux travaux étudiés. Les critères appartenant à une classe donée sont représentés par un critère englobant. Les quatre critères principaux que nous avons gardé sont:

- Les média utilisés lors de la création du résumé vidéo (Vidéo, Audio, Texte),
- la phase de prétraitement qui permet de préparer les données et de les structurer,
- la méthode d'analyse qui est le noyau de la méthodologie de création de résumés et qui définit la technique de sélection des éléments pertinents de la vidéo originale à inclure dans le résumé,
- et enfin la méthode d'évaluation et les applications des résumés résultants.

L'analyse que nous avons réalisée est récapitulée dans le tableau suivant. La première colonne comporte les références des travaux considérés, les quatre autres colonnes correspondent aux critères principaux selon lesquels, chaque travail a été analysé. Chaque ligne représente les critères spécifiques au travail de recherche correspondant à cette dernière.

Références	Média	Pré-traitement	Méthode d'analyse	Application Evaluation
HH99	Vidéo	Classif Dc_pl	A_V_C	R_S, R_D, Evl_m
GB99	Vidéo	Classif	Cls_hier	R_S, Evl_m
UF99	Vidéo	Dc_pl	M_I_P	R_S, Evl_m
DDAK99	Vidéo	Dc_pl Mpeg	App_stat	R_S, Evl_m

CSL99	Vidéo	Classif Dc_pl Mpeg	Cls_hier Th_graph	R_S, Evl_m
CI02	Vidéo	Dc_pl Mpeg	App_stat	R_S, Evl_m, Evl_h
LPE97	Vidéo Audio Texte	Classif Dc_pl	Rg_Dec	R_D, Evl_h
Duf00	Vidéo	Dc_pl	App_stat	R_S, Evl_h
GL00a	Vidéo	Classif S_éch	D_P_V App_stat	R_S, Evl_m, Evl_h
GL00b	Vidéo	Classif S_éch	App_stat	R_S, Evl_m, Evl_h
HT97	Vidéo	Dc_pl S_éch	I_Mv_P	R_D, Evl_m
ZRHM98	Vidéo	Classif Dc_pl	Rg_Dec	R_S, Evl_m
LGA+99	Vidéo	Classif Dc_pl S_éch Diff_Cr	R_Neu	R_D, Evl_m
CGP+00	Vidéo	S_éch	Alg_Gén	R_S, Evl_m
PKSW98	Vidéo	Dc_pl	Rg_Dec	R_S, R_D, Df_Appl
SK97a	Vidéo Audio Texte	Dc_pl	Rg_Dec D_P_V	R_D, Evl_h, V_Sp
HLR+99	Vidéo Audio Texte	Classif	Cls_hier	R_S, Evl_m
LPsF97	Vidéo Audio Texte	Classif Dc_pl	Rg_Dec	R_D, Evl_h
SC01	Vidéo Audio	Classif Dc_pl	App_stat Rg_Dec	R_D, Evl_m

R_D = Résumé dynamique
R_S = Résumé statique
Classif = Classification
Dc_pl = Découpage en plans
Evl_h = Evaluation humaine
Evl_m = Evaluation mathématique
Srf_aff = Surface d'affichages des images caractéristiques
Mpeg = Utilisation du flux compressé
App_stat = Approche statistique
Th_graph = Théorie des graphes
Cls_hier = Classification hiérarchique
S_éch = Sous-échantillonnage
Diff_Cr = Différentes caractérisations des images
R_Neu = Réseaux de neurones
P_M_D = Préparation manuelle de données d'apprentissage
U_Ind = Utilisation des média de façon indépendante
V_Sp = Type de vidéos spécifique (Journal télévisé)
D_P_V = Détection de la peau et des visages
Rg_Dec = Règles de décision
Alg_Gén = Algorithmes Génétiques
Df_Appl = Différentes applications = plusieurs tâches
A_V_C = Analyse de la validité de la classification
M_I_P = Définition d'une mesure d'importance de plan
I_Mv_P = Calcul de l'intensité du mouvement par plan

Après cette première présentation assez détaillée des différents travaux de recherche, du domaine de la construction des résumés vidéos, suivant différents axes, nous élaborons une synthèse générale. Cette dernière permettra de mettre en avant nos contributions par rapport aux travaux déjà effectués.

Suite à cette étude bibliographique, nous avons constaté qu'une grande partie des méthodes de construction des résumés vidéos sont menées en utilisant exclusivement les informations visuelles (Média = Vidéo) tandis qu'un certain nombre

de méthodologies combinent différents média (vidéo, audio et texte). La méthode que nous proposons est une méthode générique qui s'applique d'une part aux informations visuelles, et d'autre part aux différents types d'informations (textuelles ou audio).

Nous observons aussi que l'ensemble des méthodologies proposées comportent une phase de prétraitemement qui permet de diminuer le quantité d'informations traitées, de structurer les données en fonction de relations temporelles ou visuelles. Cette phase consiste généralement à appliquer une classification, un découpage en plan, un sous-échantillonnage ou une combinaison de ces derniers. Dans notre approche, nous appliquons un sous-échantillonnage pour diminuer le nombre d'images traités, ensuite une classification des images restantes selon leur similarité visuelle.

Lors de cette analyse, nous avons remarqué que chaque travail de recherche est distingué par sa méthode d'analyse. Cette dernière permet la selection d'images représentatives ou de sous-séquences pertinentes de la vidéos originale. La tâche de sélection est considérée comme étant un problème d'optimisation dans quelques méthodes, et elle est basée sur un ensemble de règles prédéfinies selon des besoins spécifiques dans d'autres. Les avantages et les inconvénients de ces différentes méthodes ont été présentés selon la technique utilisée dans les paragraphes précédents. Notre méthode s'inscrit dans le cadre d'optimisation. Nous construisons un résumé optimal ou sous-optimal selon un critère mathématique inspiré du problème de l'évaluation de la qualité du résumé.

En fonction de l'application pour laquelle est destiné le résumé créé, les différentes approches proposées adoptent d'une part, l'extraction de sous-séquences visuelles ou audio-visuelles de durées réduites de la séquence originale. Le rassemblement des ces sous-séquences produit un résumé dynamique qui préserve l'évolution temporelle et le dynamisme présent dans la vidéo originale. D'autre part, Elles sélectionnent un ensemble d'images représentatives du contenu global de la vidéo considérée parmis les images composant cette dernière. Cette collection d'image forme un résumé statique et permet d'avoir une idée générale du contenu visuel. Notre méthode permet la construction d'un résumé statique composé d'un ensemble d'images représentatives. Ces dernières sont considérées comme des centroïdes de segments visuels extraits de la séquence vidéo originale pour construire

un résumé dynamique.

Les méthodes de construction des résumés vidéos qui utilisent différents média traitent en général indépendamment chaque types d'information avec des méthodes appropriées, ensuite les résultats obtenus sont combinés afin de sélectionner les segments qui composeront le résumé final. Dans notre approche, un résumé multimédia est construit en adaptant le même principe de sélection sur les différents types d'information pris en considération ce qui permet une cohérence entre les média et une signification plus intuitives des résultats obtenus lors de l'évaluation de la qualité du résumé généré par notre méthode.

Malgré, le grand nombre de travaux de recherche traitant le problème de construction de résumés vidéos, aucun des groupes de recherche ne s'est attardé sur le problème de multi-vidéos où des contraintes spécifiques sont à prendre en compte. Nous nous sommes intéressés à ce problème en adaptant notre principe de construction de résumés vidéos à ce cas particulier.

Enfin, quelque soit la méthode proposée et utilisée pour la construction d'un résumé vidéo dans les projets de recherche menés jusqu'à ce jour, il n'existe malheureusement pas une méthode objective d'évaluation de la qualité de ce dernier. Ce point qui consiste à l'évaluation objective de la qualité du résumé créé est le noyau principal de notre travail. Notre méthode de création de résumés multimédia est inspirée d'une méthode d'évaluation objective par rapport à une application particulière que nous avons définie.

2.6 Résumés textuels

Plusieurs travaux ont été réalisés dans le domaine du traitement du texte et plus particulièrement avec l'intention de construire des résumés textes de différents types de documents. Les premiers travaux remontent aux années 50. Mani et Maybury [MM99] ont classifié les différentes approches proposées en quatre catégories principales: des approches classiques, des approches basées sur l'utilisation de corpus, des approches exploitant la structure du discours, et des approches utilisant des connaissances de haut niveau.

- Les approches classiques sont celles qui servent comme base fondamentale

pour des applications pratiques et modernes ainsi qu'une motivation pour de futures recherches. L'article de Luhn [Luh58] décrit une approche statistique de construction de résumés vidéos. Elle consiste en la sélection des informations importantes en fonction de la fréquence des mots. Edmundson [Edm69] a comparé l'utilisation de la fréquence des mots avec d'autres caractéristiques comportant les expressions-clés, le titre et les mots de l'entête, ainsi que l'emplacement des phrases. Pollock et Zamora [PZ75] ont présenté un programme de construction de résumés de documents de chimie destinés au «Chemical Abstractors Service». Cette méthode a procédé en utilisant les expressions-clés spécifiques au domaine de la chimie. Ces expressions ont été utilisées comme facteur positif ou négatif lors de la sélection des phrases pertinentes. Ces dernières ont été ensuite réduites en se basant sur les analyses linguistiques de faible profondeur.

- Les approches basées sur l'utilisation des corpus représentent un prolongement du travail d'Edmundson [Edm69]. Ces méthodes peuvent être utilisées pour obtenir des statistiques concernant les mots et pour décider de la combinaison appropriée de ces caractéristiques. Kupiec et al. [KPC95] ont proposé l'utilisation d'un classificateur bayesien pour extraire les phrases pertinentes. Le classificateur a été entraîné avec des vecteurs caractéristiques construits à partir des phrases du texte intégral. De même que le travail d'Edmundson, ils ont trouvé que, pour les données utilisées, l'emplacement de la phrase est la meilleure caractéristique individuelle par rapport à l'association des différentes caractéristiques considérées et la meilleure combinaison a été: l'emplacement, les expressions-clés, et la longueur de la phrase. Myaeng et Jang [MJ99] ont proposé une variante de la méthode précédente [KPC95] appliquée aux documents techniques rédigés en langue Coréenne. Ils ont trouvé que l'utilisation de la combinaison de mots clefs, l'emplacement des phrases, et la présence des mots du titre dans une phrase donne les meilleurs résultats. Aone et al. [cAGLO99] ont révélé que dans le cas de l'utilisation de mesures statistiques sur les mots en combinaison avec l'approche de Kupiec [KPC95], différentes manières de regroupement des mots selon la reconnaissance morphologique, les synonymes, les noms propres peut influer sur les performances des résumés construits.

- Les approches du troisième type étudient les modèles de structure du discours restant relativement indépendant du domaine et avec un minimum de connaissances. Boguraev et Kennedy [BK97] ont regroupé et sélectionné les mots en se basant sur une analyse syntaxique robuste ainsi que la résolution des liens entre les différents mots. Barzilay et Elhadad [BE97] ont regroupé les mots sous forme de chaînes lexicales selon la relation existant entre les mots (par exemple, deux mots sont synonymes ou hyperonymes). Ensuite, ces chaînes ont été utilisées pour la sélection des phrases. Dans d'autres travaux on a construit des modèles globaux des structures du discours en utilisant des sous-modèles basés sur les différentes relations qui peuvent exister dans un texte [Mar99] [SSWW99] [SM99].
- Les approches qui utilisent les connaissances de haut niveau, essayent de modéliser les connaissances nécessaires pour la construction des résumés en fonction du domaine et de la nature des documents utilisés [Leh81] [UU99] [KJK95] [May95]. En général, dans ce type d'approches les auteurs font l'hypothèse qu'une représentation structurée et riche en connaissances est disponible de sorte que le processus de construction puisse l'utiliser comme une entrée. Lehnert [Leh81] a décrit une technique théorique par déduction pour la construction de résumés de documents narratifs. Cette technique a été basée sur les relations structurelles dans la construction de documents narratifs.

2.7 Evaluation des résumés textes

Le problème d'évaluation de résumés textes est un problème très critique. Il est difficile de choisir la méthode ou le type d'évaluation les plus appropriés. Il existe différentes possibilités d'évaluer les performances d'un système de construction de résumés textes [MM99]. Comme exemples, nous citons, la comparaison du résumé avec le document original, la comparaison du résumé construit par le système avec un résumé créé par un être humain, et la comparaison de deux résumés construits par deux systèmes différents.

Ces diverses méthodes d'évaluation peuvent être classées en deux grandes catégories.

- La première est l'évaluation «intrinsèque», où des tiers jugent la qualité du processus de construction du résumé en se basant directement sur l'analyse du résumé lui même. Par exemple:

  - Les utilisateurs jugent la facilité de lecture du résumé [MSG97]. La mesure de la facilité d'un document texte considère la complexité du langage (comme la longueur des mots et des phrases), la présence d'anaphore, la préservation de l'environnement structuré comme les listes et les tableaux, les caractéristiques grammaticales, le style général, etc... plusieurs parmi ces mesures peuvent être calculées automatiquement.

  - Les utilisateurs vérifient si le résumé couvre les idées pertinentes traitées dans le document original [PJ93]. Ce jugement de la couverture du résumé des idées clefs abordées dans le document principal est subjectif même si les usagers sont d'accord d'avance sur ce qui constitue l'ensemble des idées pertinentes.

  - Des évaluateurs comparent le résumé construit avec un résumé qu'ils supposent comme étant idéal [Edm69] [KPC95]. Cependant la construction d'un résumé idéal reste très difficile à mettre en œuvre. Le résumé idéal peut être fourni par l'auteur même du document original, par un évaluateur à qui était demandé de créer un résumé pour le document ou un autre évaluateur à qui était demandé d'extraire les phrases pertinentes du document original. L'absence d'une perception unique du résumé idéal permet d'avoir différents résumés de bonne qualité [GAT61]. Les travaux de [Mar99] ont montré que même si différents évaluateurs ne sont pas d'accord sur l'ensemble des phrases constituant un résumé idéal, ils peuvent être d'accord sur les meilleures phrases à inclure dans le résumé parmi les phrases les plus pertinentes du document original.

- La deuxième catégorie est l'évaluation «extrinsèque», dans laquelle la qualité du résumé est déterminée en se basant sur la manière dont celui-ci influence l'accomplissement de certaines tâches. Les tâches ont inclus la

détermination humaine de l'importance d'un document par rapport à un thème donné [IE99] ainsi que l'interrogation de certaines personnes en leur demandant de répondre en se basant uniquement sur la lecture des résumés construits [AM99].

Plusieurs travaux de recherche ont traité le problème de l'évaluation de résumés textuels. Parmi ces travaux nous citons:

- Rath et al. [GAT61] qui ont présenté un travail classique montrant que différentes personnes construisent généralement des résumés différents d'un même document texte. En outre, un même utilisateur ne construira pas (d'une façon identique) un deuxième résumé du document qu'il a déjà résumé depuis plus de huit semaines. Les auteurs ont comparé cinq méthodes de construction automatique de résumés (extraction de phrases pertinentes) basées sur le calcul de fréquences ainsi que la distribution des mots. Ils ont comparé les performances de ces méthodes les unes aux autres ainsi qu'avec des sélections de phrases effectuées par six personnes différentes. Ils ont remarqué que les résumés construits automatiquement sont assez similaires tandis que ceux créés par les six utilisateurs sont assez différents. Ils ont aussi observé qu'il n'y a pas beaucoup de phrases communes entre celles extraites par les méthodes automatiques et celles sélectionnées par des utilisateurs.
- Brandow et al. [BMR95] ont décrit la conception et l'évaluation du système ANES qui a été développé à «General Electric» aux débuts des années 90. L'approche utilise la pondération des mots et des phrases en se basant sur la formule TF.IDF en éliminant les mots dont le poids est inférieur à un certain seuil. Pour résoudre le problème d'anaphore, les phrases débutant par certaines anaphores ont été éliminées. Dans le but de garder une continuité entre les phrases, ils ont sélectionné les phrases qui ne comportent pas de mots-clés mais qui font la liaison entre deux phrases comportant des mots clefs. Aussi ils ont ajouté au résumé la première (deuxième) phrase de chaque paragraphe si la deuxième (troisième) phrase contient des mots-clés. Le Système ANES a été évalué de manière intrinsèque. Des

évaluateurs ont été sollicités pour déterminer si des extraits courts de nouvelles histoires produits par le système sont acceptables en leur remettant les documents comportant l'intégralité de ces histoires. L'évaluation de la qualité de chaque résumé ont été basée sur la facilité de lecture du résumé et sa couverture des idées présentées dans les documents originaux. Les taux de satisfaction enregistrés a été de l'ordre de 68% à 78% en fonction de la longueur du résumé évalué (50-150 ou 250 mots). Les évaluateurs ont comparé ensuite le système ANES avec le système «Searchable Lead» qui est un système renommé de traitement de texte réalisé par «Mead Data Central». Ce dernier procède par la sélection de la partie initiale du document texte comme résumé du document intégral. L'évaluation de ce système a donné des performances de 87% à 96%. Ces performances sont meilleures que celles attribuées aux résumés créés par le système ANES. Il faut noter que les auteurs n'ont pas fait d'analyse statistique des résultats obtenus. De plus il n'est pas clair de savoir comment l'entête d'un document peut être efficace pour d'autres genres comme les résumés longs, les résumés personnalisés en fonction des besoins des utilisateurs, et les documents où l'information pertinente n'est pas évoquée au début du document original.

- Morris et al. [AGD92] ont présenté une méthode d'évaluation extrinsèque basée sur une tâche d'interrogation (questions réponses). Les utilisateurs ont utilisé quatre exercices de compréhension de lecture tirés des examens d'admission en école de management «Graduate Management Admission Test». Les exercices ont été sous forme de QCM, dont chacun est composé de huit questions. Une simple question correspond à cinq différentes réponses dont une seule doit être cochée. Ils ont ensuite demandé à quelques utilisateurs de faire ces exercices dans différents contextes. Les contextes étudiés ont été les suivants: les utilisateurs ont consulté le document original, les utilisateurs ont vu un résumé instructif construit manuellement par un expert avec un taux de compression de 25%, les utilisateurs ont examiné un résumé construit automatiquement, et les étudiants répondent aux questions sans avoir rien vu. Pour les résumés automatiques, ils ont testé trois types de résumés. Le premier a été construit en faisant l'extraction

aléatoire de quelques phrases du document original avec un taux de compression est de 25%. Le deuxième à un taux de compression de 20% et le troisième à un taux de 30% ont été générés par un algorithme de pondération de phrases basée sur la méthode présentée par Edmundson [Edm69] utilisant la présence de phrases fixes, les mots du titre, les mots thématiques ayant une grande fréquence, et les positions des phrases dans le paragraphe. L'analyse des résultats a montré que les performances des résumés manuel et automatique à 20% et 30% de taux de compression sont comparables aux performances du document original. Cependant les performance de l'utilisation du résumé aléatoire automatiquement construit et la non-utilisation d'un support de cours étaient très faibles par rapport aux performances du document intégral. Suite à ces résultats, les auteurs ont suggéré l'utilisation des résumés comme substitution des documents originaux pour l'accomplissement de certaines tâches.

- Firmin et al. [TM99] ont décrit un «benchmark» initial de l'évaluation des systèmes de construction de résumés vidéos TIPSTER menée en 1997. Cette évaluation est particulièrement éminente parce que c'est la première évaluation sur une grande échelle menée par des utilisateurs n'ayant pas développé de systèmes. C'est une évaluation extrinsèque basée sur une tâche qui permet de mesurer l'impact du résumé sur la performance de temps et la précision de l'estimation de la pertinence. L'hypothèse de l'expérience est que l'utilisation du résumé permet la préservation des temps d'estimation de la pertinence dans l'accomplissement de certaines tâches. Dans la première tâche ad hoc, les utilisateurs ont été sollicités pour juger si un texte (document original ou résumé) est pertinent ou non par rapport à une requête donnée. Dans la deuxième tâche de classification, un texte (résumé ou document intégral) a été jugé comme appartenant à une classe parmi cinq classes de sujets mutuellement exclusives ou ne faisant partie d'aucune des classes considérées. La rigueur des utilisateurs dans la détermination de pertinence d'un document par rapport à une requête ou une classe a été estimée par le jugement de la pertinence grâce à une vérité de terrain utilisée dans les conférences TREC [HV98] qui sont les évaluations standards des systèmes de recherche d'information menées par le gouvernement

américain. Pour chaque tâche deux types de résumés ont été utilisés: des résumés de longueur fixe (10% du document original), des résumés ayant la meilleure longueur (sans aucune restriction). Afin de comparer ils ont aussi utilisé un résumé de base comportant les premiers 10 % du document intégral. En analysant les résultats des expériences, ils ont observé que les temps d'estimation de pertinence en utilisant les résumés sont plus courts que ceux enregistrés en utilisant les documents intégraux. Ils ont constaté aussi qu'il n'y a pas de différence significative entre les performances des divers résumés utilisés.

Il existe un certain nombre de problèmes fondamentaux qui ressort lors de la conception des évaluations des systèmes de construction des résumés textes, Mani et al. [MM99] ont cité:

- La création de tâches pour l'évaluation extrinsèque qui modélisent des situations du monde réel en essayant de répondre à des besoins concrets. Il est souvent difficile d'automatiser des tâches réelles en réalisant des expériences contrôlées. Idéalement, les tâches créées sont des tâches qui sont pénibles et lourdes pour les utilisateurs, cependant les systèmes automatiques peuvent les accomplir avec une grande vitesse.
- La préparation d'un ensemble d'instructions que les personnes suivront lors de la création des résumés utilisés dans l'évaluation intrinsèque. Comme par exemple, construire un résumé soit en fonction de ce qu'ils ont compris de la lecture du document original, soit en faisant une extraction des phrases ou des sous-phrases. Dans les deux cas préciser le taux de compression par rapport au document intégral. La qualité ainsi que la nature du résumé résultant dépend directement des ces instructions.
- La vérification que la tâche n'est pas biaisée par rapport à la technologie de la construction ou le genre et la longueur du document à résumer.
- Faire la conception d'une expérience qui traite assez de données afin d'obtenir des résultats sur lesquels des opérations statistiques peuvent être menées. Souvent, les ressources nécessaires pour mettre en œuvre ce type d'expériences sont excessives. Parfois, il est plus intéressant de choisir une évaluation réalisée par des experts humains que par des expériences automatiques.

- Trouver une mesure appropriée pour le calcul de l'efficacité et la précision du processus de construction de résumé.

2.8 Evaluation des résumés vidéos

Peu de travaux de recherches ont été consacrés au problème de l'évaluation des résumés vidéo. Jusque là, il n'y a pas eu de benchmarks pour évaluer les différents systèmes proposés dans ce domaine de recherche. Quelques auteurs évaluent la qualité de leurs résumés en utilisant des mesures statistiques. Les performances calculées sont souvent incompréhensibles par rapport aux critères du jugement humain. D'autres impliquent un certain nombre d'utilisateurs réels dans la phase d'évaluation. Cette intervention est très utile et significative mais par contre très difficile à mettre en œuvre.

Les deux exemples suivants décrivent une évaluation basée sur l'intervention humaine. La première est intrinsèque, cependant la deuxième est extrinsèque.

- D. Diklic et al. [DPD98] ont décrit des nouveaux algorithmes de sélection automatique d'images-clés représentatives basés sur le contenu des scènes. L'ensemble des algorithmes proposés peut être divisé en trois groupes ou catégories, la première consiste à calculer l'image moyenne de toutes les images composant la vidéo, puis calculer les distances respectives de ces dernières par rapport à cette image moyenne. Dans la deuxième catégorie d'algorithmes, Les auteurs ont calculé les différences des histogrammes des images successives pour sélectionner les images-clés. Le troisième groupe utilise des mesures statistiques (comme la variance d'histogramme) de chaque image afin de retrouver les plus représentatives. Pour chaque catégorie, sont développées deux ou trois variantes d'algorithmes en modifiant l'espace de couleur ou l'image est représentée, à titre d'indication par l'espace RGB, l'espace LUV ou le niveau de gris. Les résumés construits par l'utilisation des neufs algorithmes résultant de la combinaison des 3 méthodes proposées avec les 3 espaces de couleurs ont été présentés à quelques utilisateurs réels pour faire une évaluation intrinsèque. Différents résumés d'une même vidéo créés par l'une des méthodes sont affichés sur une page Web. L'utilisateur

attribue une note entre 1 et 5 à chaque résumé selon son jugement personnel de la qualité par rapport aux autres et à la vidéo originale.

- Ding et al. [DMT97] ont proposé une méthode de construction d'un résumé vidéo en sélectionnant un nombre d'images-clés par l'utilisation d'un sous-échantillonnage uniforme (une image prise par chaque intervalle de temps fixe). Ensuite, les images sélectionnées ont été visualisées dans l'ordre séquentiel de leur position dans la vidéo. Les auteurs ont testé lors de l'affichage différentes fréquences d'images par seconde. Afin d'évaluer la qualité des résumés résultants, ils ont adopté une méthode d'évaluation extrinsèque. Ils ont défini deux tâches, puis ils ont demandé à quelques personnes (20) d'accomplir ces tâches en ayant uniquement connaissance du résumé présenté.

  - La première tâche est d'expliquer ce qu'ils ont retenu et compris de celui-ci. Leur compréhension du contenu est ensuite mise en correspondance avec le contenu du document original.

  - La deuxième tâche consiste à reconnaître les objets présents dans les résumés présentés. Pour ceci, une liste d'objets dont la moitié correspond à des objets non présents dans la vidéo originale leur avait été distribuée. En visualisant, un résumé donné, les évaluateurs doivent sélectionner parmi les objets de la liste, ceux présents dans le résumé.

  Les performances des différents utilisateurs ont été ensuite comparées pour définir le meilleur résumé (meilleure fréquence d'affichage).

2.9 Conclusion

Dans ce chapitre, nous avons présenté une revue générale des approches de construction et d'évaluation de résumés vidéos et texte. Le but de cette étude est de situer notre travail par rapport à quelques travaux déjà effectués dans ce domaine. Nous avons observé que malgré le grand nombre de méthodes proposées pour la construction de résumés vidéos, il n'y a pas eu vraiment de méthodes d'évaluation objective de la qualité de ces résumés par rapport aux besoins réels

des utilisateurs sauf dans quelques travaux qui se basent sur une intervention directe des utilisateurs. A travers cette étude, nous avons aussi remarqué qu'il n'y a pas eu de travaux consacrés au cas des résumés multi-vidéos. Les approches de construction de résumés vidéos basées sur plusieurs média analysent le contenu de chaque flux indépendamment de l'autre. Ils sélectionnent les segments qui composeront le résumé multimédia, soit en utilisant un seul média (vidéo ou audio ou texte) puis en combinant les segments sélectionnés de ce flux avec les segments correspondants des autres média, soit en utilisant un ensemble de critères pour combiner les différents flux lors du processus de sélection. Notre idée intuitive inspirée du problème de l'évaluation, nous a amené à proposons une méthode de construction de résumés optimaux par rapport à une tâche que nous définissons au préalable. Cette tâche qui peut être accomplie par des utilisateurs réels lors d'une évaluation extrinsèque. Elle est automatisable, ce qui nous permet une évaluation objective de la qualité des résumés construits sans intervention directe des utilisateurs. Pour obtenir ce résultat le comportement de l'utilisateur réel est simulé selon un nombre d'hypothèses, par le système. Cette méthode de construction peut être utilisée pour la construction de résumés vidéos en utilisant différents média indépendamment ou combinés. Enfin, notre méthode peut être adaptée au cas des résumés multi-vidéos.

Chapitre 3

Principe de Reconnaissance Maximale

Dans ce chapitre, nous posons le problème de la construction de résumés vidéos optimaux, précisons la notion d'optimalité pour ce problème et présentons une approche de construction originale tout en soulignant les raisons qui nous ont conduits à l'adopter. Cette approche est basée sur un principe général dit de Reconnaissance Maximale (PRM). Le concept de reconnaissance maximale que nous présentons ici d'une manière générale sera adapté aux diverses situations pratiques étudiées dans la suite de ce travail.

3.1 Introduction

Nous proposons dans ce travail de recherche un outil capable d'aider les utilisateurs à consulter le contenu essentiel des vidéos selon leur temps de disponibilité: nous nous penchons sur le problème de la construction de résumés vidéos. Notre objectif va consister à définir et à construire ce nouvel outil de création de résumés vidéos «optimaux».

En théorie, un résumé «optimal» serait celui qui représente le mieux possible le contenu du document vidéo original par rapport à certains critères. En pratique et dans un contexte général, l'optimum dépend des utilisateurs, de leur besoins, et du type de document audiovisuel considéré. Par exemple, certaines personnes se souviennent mieux d'un film vu longtemps auparavant que du journal télévisé de la veille; deux personnes peuvent se rappeller d'une émission ou

d'un documentaire, sans pour cela en retenir les mêmes détails; deux téléspectateurs peuvent retenir les mêmes passages d'un match de football, mais en oublier le reste, etc... Par ailleurs, le taux d'informations mémorisées par un individu dépend de plusieurs facteurs, comme par exemple son attention, sa motivation, ou sa concentration. Ce que se rappelle un individu donné reste très personnel. Une chose commune à l'ensemble des individus est qu'ils se rappellent en général ce qu'ils considèrent comme des éléments «importants»; en d'autres termes, chaque personne résume et structure à sa manière le document multimédia, et ne garde en mémoire qu'une fraction représentative de ce dernier.

Dans le but d'unifier les avis des différents utilisateurs, le résumé peut être lié à une tâche donnée; ainsi, par exemple, le résumé optimal sera celui qui permet à l'ensemble des usagers de bien accomplir cette tâche. De ce fait, la construction d'un résumé devient une opération objectivisée, et dont le but est de permettre la meilleure réalisation possible de la tâche associée. Une fois qu'un tel résumé est construit, le problème qui se pose est d'en évaluer la «qualité». Cette question est discutée ci-dessous.

Imaginons que nous disposons d'un groupe de volontaires pour nous aider à réaliser l'expérience suivante. Nous demandons à ces usagers de regarder une vidéo, de lire un document, ou plus généralement de prendre connaissance d'un document multimédia que nous mettons à leur disposition. Ensuite, nous sollicitons chacun d'eux pour qu'il fasse un résumé représentatif du document qu'il a observé. Une fois que les résumés sont construits par ces utilisateurs, et sachant que nous disposons des documents originaux, nous essayons d'évaluer la qualité de chacun des résumés créés. L'évaluation peut être faite de plusieurs manières, et selon différents critères. Ces diverses méthodes d'évaluation peuvent être classées en deux grandes catégories [MM99]. La première est l'évaluation «intrinsèque», où des tiers jugent la qualité du processus de construction du résumé en se basant directement sur l'analyse du résumé lui même. La deuxième catégorie est l'évaluation «extrinsèque», dans laquelle la qualité du résumé est déterminée en se basant sur la manière dont celui-ci influence l'accomplissement de certaines tâches.

Afin de préciser ces notions, nous pouvons citer comme exemples d'évaluation intrinsèque les procédures suivantes:

- Demander à un jury de spécialistes de juger directement la qualité de chaque résumé en donnant par exemple des notes, ou en faisant des observations.
- Etudier les caractéristiques de chaque résumé, comme sa taille, sa clarté, sa couverture, sa diversité, son attraction, etc... Ensuite, combiner ces critères pour établir une évaluation générale.
- Comparer les résumés obtenus avec un résumé supposé idéal.

De manière analogue, les méthodes suivantes peuvent être classées dans les procédures d'évaluation extrinsèque:

- Répondre à un ensemble de questions concernant le document original en ayant uniquement connaissance de son résumé.
- Déterminer l'importance d'un document par rapport à un thème donné.
- Identifier si des extraits donnés proviennent d'un document original en n'ayant connaissance que de son résumé.
- Classer des passages textuels ou visuels pris parmi plusieurs documents, de telle sorte que les passages qui proviennent d'un même document appartiendront à la même classe.

Les méthodes d'évaluation intrinsèques présentées ci-dessus ont en commun le fait qu'elles mettent en œuvre des mécanismes d'évaluation subjective. Il semble donc difficile de mettre en place pour ces méthodes des systèmes informatiques créant automatiquement les résumés optimaux correspondants. A l'inverse, les méthodes extrinsèques se soumettent plus facilement à une telle automatisation en raison des critères plus objectifs qui les constituent. Par conséquent, c'est vers ces méthodes que nos recherches vont se tourner.

Dans l'expérience imaginaire décrite plus haut, envisageons maintenant une évaluation extrinsèque qui consiste à définir une ou plusieurs tâches pour évaluer la qualité des résumés construits par les différents volontaires. Afin d'évaluer la qualité de chaque résumé, nous devrions donc déterminer dans quelle mesure chacun de ces résumés permet à des utilisateurs (n'ayant pas une connaissance préalable des documents multimédias originaux) de réaliser les tâches associées. La performance avec laquelle est réalisée une tâche donnée sera en quelque sorte un miroir reflétant la qualité du résumé, qui a aidé l'usager pour l'accomplissement de cette tâche.

Pour identifier le meilleur résumé qui sera associé à une tâche donnée, nous devrions donc comparer les performances obtenues par certains utilisateurs lors de l'accomplissement de cette tâche (ces utilisateurs ayant connaissance d'un résumé parmi ceux construits par des volontaires différents, qui eux ont observé le même document original). Finalement, nous sélectionnerons le résumé qui permet d'avoir les meilleures performances. Pour que la sélection d'un bon résumé soit fiable, il faut, pour chaque tâche, comparer l'utilisation de plusieurs résumés créés par différents utilisateurs, et demander à des tiers de la réaliser. Ces derniers ne doivent pas avoir consulté le document original, ni d'autres résumés, afin que leur comportement ne soit pas influencé. En effet, un utilisateur qui a déjà participé à l'évaluation d'un premier résumé ne peut être fiable pour évaluer un autre résumé du même document, vu qu'il a déjà une certaine connaissance du contenu correspondant à ce document. Le nombre d'utilisateurs nécessaire pour créer et évaluer des résumés quasi optimaux de plusieurs documents par rapport à différentes tâches peut ainsi devenir arbitrairement grand. Ceci rend la généralisation de cette expérience très coûteuse: il devient difficile de réaliser en pratique une telle évaluation extrinsèque pour un grand nombre de documents multimédia. Afin de mettre en œuvre une procédure d'évaluation réalisable, nous proposons dans ce travail d'*automatiser* le processus de construction et d'évaluation de résumés optimaux de documents multimédia. Ce processus sera automatisé par rapport à une tâche réaliste, pour laquelle le comportement des utilisateurs réels lors de la réalisation de cette tâche pourra être simulé par notre système. Dans ce cadre, nous choisirons la tâche réaliste suivante: «Reconnaître si un passage ou un extrait d'un document multimédia provient du document original du résumé créé ou non». Ensuite, nous engendrerons un processus automatique de construction et d'évaluation de résumés multimédia quasi-optimaux par rapport à cette tâche, et par rapport au modèle d'attitude généré par la simulation du comportement des utilisateurs réels.

3.2 Idée Intuitive

En raison du grand nombre de chaînes télévisuelles existantes, le téléspectateur peut être amené à zapper fréquemment afin de repérer le programme qui répond

le mieux à ses attentes. Supposons que chaque fois que le téléspectateur change de canal et se retrouve au milieu d'un nouveau programme télévisé, il le regarde quelques secondes avant de tenter de le reconnaître. En visualisant une image instantanée ou un court extrait d'un programme en cours, l'utilisateur pourra reconnaître ce programme rapidement et avec certitude si nous faisons l'hypothèse qu'il se souvient de la totalité des programmes qu'il a déjà regardé. Cette hypothèse est bien sûr irréaliste. Notre but est de permettre aux utilisateurs de reconnaître efficacement un ou plusieurs programmes, sans qu'ils les aient auparavant vus ou mémorisés. Pour ceci, nous allons représenter chaque programme diffusé par un résumé constitué de quelques images à mémoriser. Ce résumé doit permettre à chaque utilisateur de conserver, vis-à-vis du programme qu'il observe, une attitude aussi proche que possible de celle qu'il aurait s'il avait déjà vu le programme en entier. Le problème consiste donc à trouver les éléments de la vidéo originale qui permettent le mieux d'identifier le programme diffusé. Précisément, nous entendons par «identifier» le fait de se rappeler avec certitude avoir déjà vu le programme en question (si le téléspectateur a déjà vu ce programme, il a gardé en mémoire certains éléments, comme par exemple des images, des séquences, des mots, etc...). Pour résumer, nous voulons construire le meilleur résumé dans le sens où, bien que le téléspectateur n'ait jamais vu en entier le programme devant lui, il trouve la correspondance la plus forte possible entre les éléments du résumé et ceux diffusés sur le canal, de sorte qu'il reconnaisse le programme en question.

Il est clair que les images sont riches en informations, et aident par conséquent la personne à l'identification. Pour préciser et formaliser notre démarche, nous allons considérer chaque programme comme étant une séquence vidéo SV composée d'une succession d'images I, sans prendre en compte à ce niveau les autres médias associés (i.e. audio et texte). En zappant, l'utilisateur U pointe une image aléatoire qui correspond à un instant t de cette vidéo SV. A partir de cet instant, il va regarder un extrait d'une certaine durée, sur lequel il va se baser pour identifier le programme.

Afin d'éclaircir notre idée, nous considérons deux cas. Dans le premier, l'observateur se base sur une image unique, correspondant à l'instant t (celle sur laquelle il tombe en zappant) pour essayer de reconnaître la vidéo SV. Dans le deuxième cas, il observe un extrait composé d'une série de plusieurs images

débutant à l'instant t, avant de tenter d'identifier la séquence vidéo.

- Considérons le cas où l'observateur se base sur une seule image.
  En premier lieu, faisons l'hypothèse que nous désirons construire un résumé $R1$ composé d'une seule image également, résumé qui doit représenter le contenu d'un programme télévisé de la meilleure façon possible (i.e. telle que l'utilisateur le mémorisera, en substitution de cette séquence vidéo qu'il n'a pas vue). Cette image sera utilisée par le téléspectateur pour reconnaître le programme dont elle est extraite. En zappant, l'utilisateur se retrouve à un instant aléatoire du programme correspondant à une image quelconque. Pour augmenter la probabilité que l'utilisateur reconnaisse l'image diffusée en la comparant avec celle qu'il a mémorisée, nous devons sélectionner l'image qui a le plus de chance de correspondre à n'importe quel instant t du programme télévisé. Cette image, que nous considérons comme la plus représentative du contenu de la vidéo du point de vue de la reconnaissance du programme, est appelée l'image la plus «fréquente» du programme pour lequel nous voulons construire un résumé.
  Afin de définir la «fréquence» d'une image I, considérons l'organisation générale des documents (audio)visuels. Notons tout d'abord qu'au sens mathématique du terme (relatif au nombre d'occurrences), chaque image n'apparaît en général qu'une seule fois dans un document visuel, et que par conséquent chaque image a la même fréquence (1 divisé par le nombre total d'images). Nous allons chercher ici à définir la fréquence différemment, en se basant sur un critère de *similarité*, plutôt que de stricte égalité. Un programme visuel est construit en regroupant une suite de scènes, chaque scène étant composée d'un certain nombre de plans, contenant eux-mêmes une succession d'images. Les mêmes images ainsi que les mêmes plans peuvent parfois être réutilisés dans différentes scènes. Par ailleurs, les images incluses dans le même plan présentent généralement des relations spatio-temporelles (i.e. des contenus visuels très semblables). Nous définirons alors la «fréquence» d'une image I comme le nombre total d'images contenues dans la vidéo présentant une similarité visuelle avec cette dernière. Plusieurs méthodes pour caractériser formellement la similarité visuelle seront présentées de façon détaillée dans le chapitre suivant.

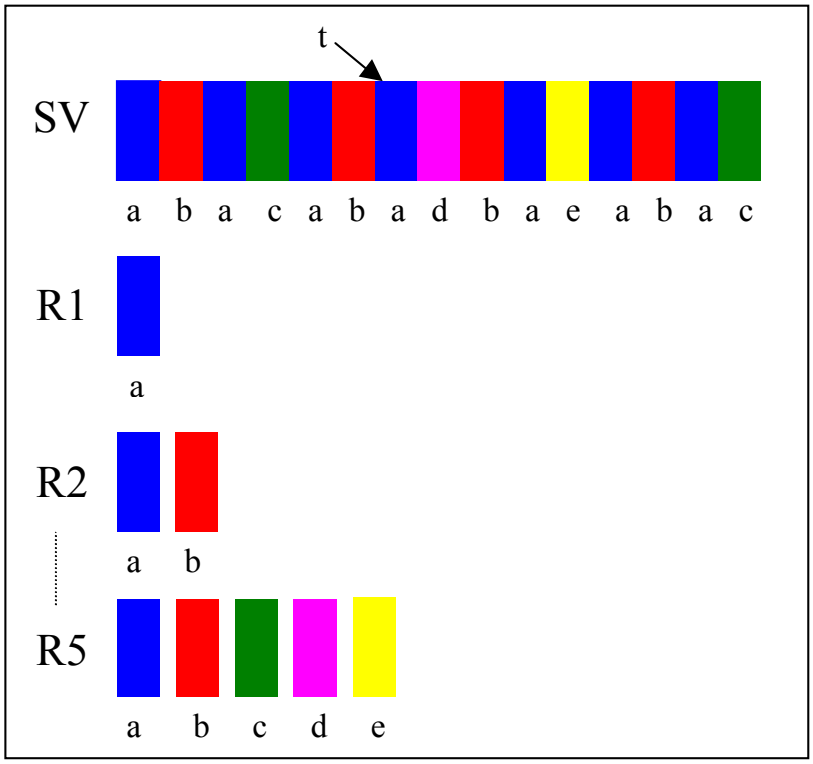

Figure 3.1: Principe de construction de résumés: cas où l'utilisateur devine à l'aide d'une seule image.

t est l'instant aléatoire auquel il commence à visionner la séquence vidéo originale (SV).

Les résumés (R1, R2,...,R5) présentés sont optimaux dans le sens où la fréquence des images qui les constituent sont classées par ordre décroissant, en partant de l'image la plus probable.

Résumons ce cas de figure: nous désirons construire un résumé composé d'une seule image $R1$ caractéristique du contenu d'un programme vidéo, permettant aux téléspectateurs de reconnaître ce programme en ayant uniquement connaissance de $R1$, et en visionnant une seule image aléatoirement choisie de ce programme; pour ce faire, nous sélectionnons l'image la plus fréquente (au sens défini ci-dessus) dans la vidéo originale.

En second lieu, afin d'augmenter les chances des utilisateurs de reconnaître le programme, nous pouvons aussi rajouter une deuxième image au résumé, telle que celle-ci soit classée en deuxième place par rapport à sa fréquence.

Nous obtiendrions ainsi un résumé $R2$. De même, si nous souhaitions construire un résumé de n éléments Rn, nous trierions les images en fonction de leurs fréquences, puis nous sélectionnerions les n premières images pour les insérer dans le résumé Rn. Ceci nous permet de maximiser la probabilité que l'image correspondant à l'instant t soit similaire à au moins une autre image du résumé de la séquence vidéo SV que le téléspectateur a mémorisé. Le résumé facilite la tâche *Identifier le programme*; il indique le contenu général de la vidéo en donnant une idée globale des sujets traités. La figure 3.1 illustre les idées précédentes. La bande schématise un programme ou une séquence vidéo SV composée d'une succession d'images. Chaque motif symbolise une succession d'images que nous supposons similaires. La flèche indique la correspondance entre l'instant t et l'image que va visionner l'utilisateur U. Un résumé composé d'un seul élément, incluant une image choisie aléatoirement dans la séquence vidéo SV, n'est probablement pas celui qui aidera le mieux les téléspectateurs pour reconnaître un programme. Dans cet exemple, le meilleur résumé à une image doit comporter l'image symbolisée par la lettre «a», car c'est la plus fréquente dans la séquence originale SV. De même, les résumés $R2$ et $R5$ comportent respectivement les deux et les cinq images les plus fréquentes dans SV.

Nous avons jusqu'à maintenant fait l'hypothèse que l'utilisateur ne se sert que de la première image visionnée comme élément de comparaison avec la (ou les) image(s) du résumé mémorisée(s) pour la reconnaissance du programme. Imaginons que le téléspectateur n'arrive pas à identifier le programme au vu d'une seule image, mais seulement après avoir visualisé un extrait d'une certaine durée comportant plusieurs images. Dans ce cas, chaque image visionnée sera l'objet d'une comparaison avec l'ensemble des images constituant le résumé que le téléspectateur garde en mémoire. Le résumé qui comporte les images les plus fréquentes de la vidéo d'origine sera-t-il alors optimal, c'est à dire celui qui permettra à l'utilisateur de reconnaître la plupart du temps le programme, quel que soit l'instant t auquel il commence à le regarder ? Nous traitons cette question dans le point suivant.

- Dans le cas où l'observateur dispose d'une suite d'images pour identifier

le programme, la distribution des images dans la séquence originale influe directement sur le choix des images rajoutées au résumé. Afin de mieux cerner cette idée, prenons l'exemple de la figure 3.2. Supposons d'abord que la taille de l'extrait soit égale à deux images. Le résumé optimal de taille un sera toujours celui qui contient l'image la plus fréquente dans la séquence originale (qui correspond dans cette exemple à la lettre «i»). Ceci est dû au fait que même si la taille des extraits visualisés par le téléspectateur est égale à deux ou plus, ces extraits contiendront plus souvent l'image caractérisée par la lettre «i» que n'importe quelle autre image. Autrement dit, si nous considérons l'ensemble de tous les extraits possibles, l'image la plus fréquente sera par définition celle qui appartiendra au plus grand nombre d'extraits par rapport aux autres images constituant la vidéo. La présence de cette image dans le résumé permet donc aux utilisateurs qui l'ont sauvegardée comme substitution de la vidéo originale de reconnaître le programme diffusé, si l'extrait visualisé contient cette image (ou une image similaire) quelle que soit la durée de ce dernier. Cette méthode optimise la performance de notre résumé par rapport à la tâche d'identification.

Supposons maintenant que nous désirions incrémenter la taille du résumé en ajoutant une deuxième image, afin d'augmenter le taux de reconnaissance par l'utilisateur (i.e, le nombre d'extraits menant à une reconnaissance correcte du document original). Pour ce problème, nous devons sélectionner une image présente dans un maximum d'extraits ne contenant pas déjà la première image. Nous remarquons que bien que les images symbolisées par les les lettres «i» et «j» soient les plus fréquentes, il y a un lien de voisinage étroit entre elles. Chaque fois que la première image apparaît dans la séquence vidéo, la deuxième la suit presque systématiquement: la plupart des extraits comportant l'image qui correspond à la lettre «i» comporteront aussi celle qui correspond à la lettre «j». De ce fait, la deuxième image engendre une redondance, et n'apporte pas vraiment une information complémentaire de la première. Dans l'exemple présenté, il semble par conséquent plus judicieux de remplacer cette deuxième image par l'image correspondant à la lettre «k», qui est relativement fréquente, et *indépendante* de la première image sélectionnée. De même, chaque fois que nous

souhaiterons augmenter la taille du résumé en rajoutant une nouvelle image, nous prendrons en compte le lien de voisinage entre les différentes images de la séquence vidéo originale, par rapport à la taille de l'extrait visualisé.

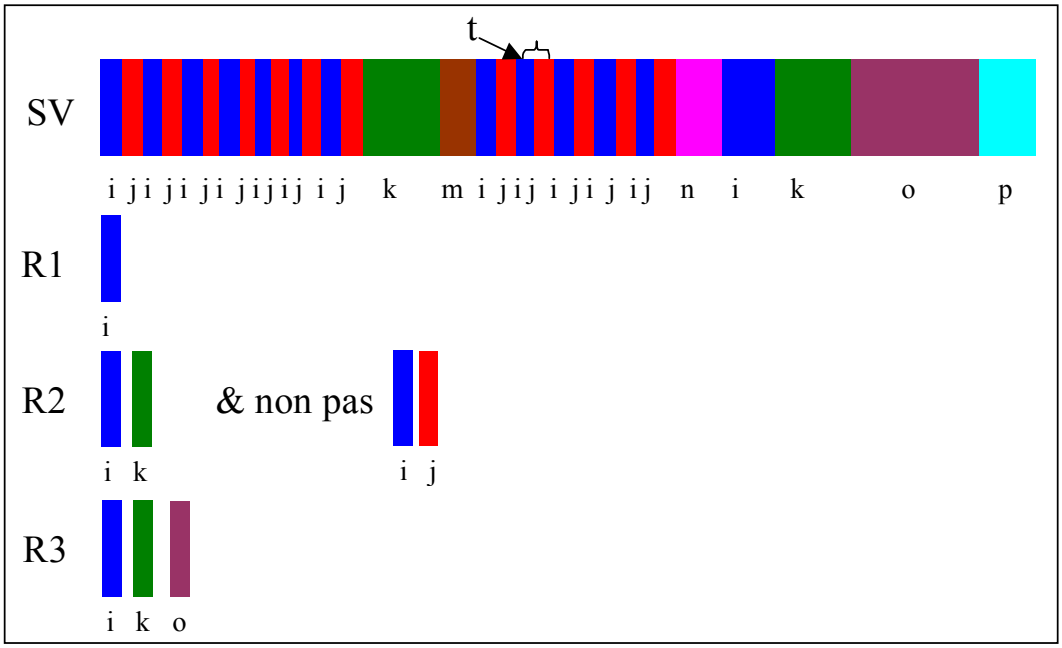

Figure 3.2: Principe de construction de résumés: cas où l'utilisateur devine à l'aide d'un extrait d'images.

t est l'instant aléatoire auquel il commence à visionner la séquence vidéo originale (SV).

Les résumés (R1, R2, R3) présentés sont optimaux dans le sens où les recouvrements des images sont complémentaires à travers la séquence (SV).

Le but de cette méthode est de maximiser la «couverture» des images du résumé. La couverture d'une image est égale au nombre des extraits de taille prédéfinie contenant au moins une occurence de cette image ou une autre qui lui est similaire. Les extraits contenant ces images doivent être complémentaires, et pas appariés ; cette méthode de construction du résumé à plusieurs images augmente le nombre des extraits permettant à l'utilisateur de reconnaître le programme diffusé au moyen de ce seul résumé.

Le résumé optimal peut alors être construit en procédant d'abord à une énumération exhaustive de tous les ensembles possibles de taille prédéfinie,

composés d'images de la vidéo originale. On additionne ensuite les couvertures relatives des images appartenant à chaque ensemble, et on garde finalement l'ensemble ayant la couverture maximale.

En se basant sur cette idée, nous pouvons imaginer une application particulière qui consiste à montrer un résumé vidéo d'un document multimédia à un utilisateur, puis à lui présenter un extrait tiré aléatoirement de la vidéo en question. On lui demande finalement de deviner si cet extrait était tiré ou non de la vidéo dont le résumé lui a été présenté. Dans les chapitres suivants, nous présenterons plusieurs méthodes pour construire les résumés optimaux par rapport à cette tâche de reconnaissance maximale, en fonction du ou des médias pris en considération.

Maintenant que nous avons présenté cette idée de façon intuitive, récapitulons et formalisons notre Principe de Reconnaissance Maximale.

3.3 Reconnaisance Maximale

Le résumé est un sous-ensemble du document original. Chaque sous-ensemble tiré du document original constitue un résumé potentiel, dont la qualité est aléatoire (elle peut être bonne ou mauvaise, en fonction de la tâche considérée).

Notre approche de création et d'évaluation de résumés vidéos est basée sur la tâche de reconnaissance. Dans ce cas, l'utilisateur est sollicité pour décider si le court extrait qui lui est présenté provient ou non de la séquence audio-visuelle originale dont il ne connaît que le résumé. La performance de l'utilisateur est définie comme le pourcentage de décisions correctes lorsqu'on considère tous les extraits possibles de la séquence originale. En d'autres termes, c'est le nombre d'extraits pour lesquels l'utilisateur devine correctement s'ils proviennent de la séquence vidéo correspondante au résumé qui lui est présenté, par rapport au nombre total d'extraits possibles. Nous dénommons cette tâche la Tâche de Reconnaissance (TR), et nous sélectionnons le résumé qui permet à l'utilisateur d'identifier et de reconnaître le plus grand nombre possible d'extraits. Le résumé optimal par rapport à la tâche définie est construit en fonction d'un Principe de Reconnaissance Maximale (PRM).

Ce principe peut être formalisé comme suit:

- Soit D un document multimédia (une séquence audio-visuelle, un texte, un document audio, etc...),
- Soit R un résumé (un sous-ensemble) de D,
- Soit E un extrait aléatoire (un sous-ensemble continu) du document D,

Nous faisons l'hypothèse que l'utilisateur U dispose d'une règle de décision $d(E, R)$ qui lui permet de décider si un extrait E provient du même document que le résumé ou non ($d = 1$ pour oui , $d = 0$ pour non),

La performance perf dans la TR est donc la valeur moyenne de $d(E, R)$ à travers tous les extraits possibles ξ tirés du document D:

$$\text{perf}(R) = \underset{\xi}{moyenne}\ d(E, R) = \frac{1}{|\xi|} \sum_{E \in \xi} d(E, R) \tag{3.1}$$

Avec cette définition, le meilleur résumé, $\hat{R}$ par rapport au PRM est:

$$\hat{R} \ = \ \arg\max_{R} \text{perf}(R) \tag{3.2}$$

Notons que cette approche dépend de la définition de la règle de décision $d(E, R)$. Nous présenterons dans les chapitres suivants plusieurs exemples de règles de décision en fonction de l'application, qui utiliseront soit la vidéo, soit le texte, soit les deux.

L'avantage de cette approche réside dans le fait que la performance du résumé n'est pas seulement un nombre abstrait mais possède une interprétation intuitive et directe. La performance est la proportion des extraits de la vidéo originale reconnus par un utilisateur n'ayant connaissance que du résumé. Un bon résumé va, bien entendu, permettre à l'utilisateur d'identifier un plus grand nombre d'extraits qu'un mauvais résumé. Cette approche fournit donc une mesure d'évaluation objective et significative de la qualité du résumé créé.

Le critère de performance étant défini, les chapitres suivants vont étudier précisément les *procédures de construction* du meilleur résumé possible. Si ces procédures s'avèreront trop coûteuses en temps de calcul, nous allons chercher à concevoir des méthodes de construction sous-optimales, mais moins complexes.

En pratique, certaines contraintes pour le calcul de la moyenne et du résumé optimal en général devront être prises en compte. Nous poserons comme paramètre la *durée des extraits* qui seront considérés lors des expériences, et nous fixerons a priori la *durée du résumé* (en terme de nombre d'images représentatives). Ceci permet de répondre au mieux aux besoins des utilisateurs, dont le temps de disponibilité et les exigences diffèrent de l'un à l'autre. Ces deux valeurs seront ainsi des paramètres, et nous étudierons leur influence dans les divers processus de construction automatique de résumés proposés dans la suite.

3.4 Conclusion

Après avoir présenté, dans ce chapitre, les diverses problématiques liées à la construction de résumés vidéos, nous avons proposé une nouvelle approche de création et d'évaluation automatique de résumés vidéos. C'est dans ce courant d'idées que nous avons exposé notre principe de reconnaissance maximale (PRM) pour la construction de résumés basée sur une tâche de reconnaissance (TR). Cette dernière peut être utilisée pour plusieurs applications, et nous montrerons plus tard qu'elle s'applique à différents types de média. Le principe de construction de tels résumés est générique ; nous l'avons décrit en faisant apparaître l'importance d'une règle de décision. Dans la suite de ce travail, nous allons nous pencher sur plusieurs exemples de règles de décision, applicables à différents média (vidéo, texte, audio).

Chapitre 4

Construction de Résumés Vidéos

Dans ce chapitre, nous détaillons notre approche pour la création et l'évaluation automatique de résumés vidéo. Pour ceci, nous adaptons le principe de reconnaissance maximale (PRM) défini dans le chapitre précédent en utilisant à ce stade uniquement l'information visuelle. Nous formulons cette approche d'une manière mathématique, et nous montrons comment cette dernière nous permet de construire un résumé quasi optimal par rapport à la tâche définie.

4.1 Introduction

Plusieurs travaux ont déjà porté sur le problème de la construction automatique du résumé d'une seule vidéo à la fois [DDK00] [IL96] [SK97b] [DKAM96] . L'approche de base consiste à opérer une classification des images ou des segments vidéos, parfois après avoir effectué un découpage en plans. Ensuite, un critère de nature mathématique permet de sélectionner les segments les plus représentatifs pour constituer le résumé. Parmi ces critères, on trouve l'utilisation des fréquences d'apparition, mais aussi parfois des contraintes sur la disposition temporelle, des notions d'abscisse curviligne lorsque la vidéo est considérée comme une courbe, ou encore l'utilisation d'une décomposition en composantes principales. Quelques rares travaux combinent plusieurs sources d'information: la vidéo, ainsi que des éléments d'analyse du signal de la parole (bruits, musique,

identification des locuteurs, reconnaissance de la parole) et même parfois le contenu textuel des sous-titres [LPE97] [PLKE98] . Ce sont alors des règles particulières qui gèrent la combinaison de ces éléments pour l'identification des moments importants. La construction automatique de bandes d'annonces de films est un exemple d'application de ces travaux [SK97a]. Par ailleurs, la plupart des approches actuelles souffrent d'un problème critique: l'évaluation de la qualité du résumé. Il est donc très délicat et très difficile d'apporter un jugement sur la performance et la qualité du résumé résultant. Même si cette dernière est calculée en utilisant un critère et une mesure mathématique, l'interprétation et la compréhension du sens restent très complexes.

Le début de ce chapitre présente l'application de notre principe de reconnaissance maximale (PRM) à l'information visuelle. Nous y présentons d'une manière formelle notre méthode de construction et d'évaluation de résumé visuel d'une seule vidéo. Ensuite nous expliquons la fonction de similarité des images que nous avons adoptée afin de sélectionner une représentation des images et une distance de mesure de similarité adéquate et raisonnable pour l'accomplissement de notre expérimentation. Enfin, nous présentons et commentons (section 4) les différents résultats obtenus.

4.2 Résumés Vidéos

Tandis qu'il existe diverses méthodes de création de résumés vidéos [YMH01c] [MT96], la visualisation des résumés résultants se fait souvent selon deux approches: le résumé est soit représenté sous une forme «statique», soit «dynamique»:

- Un résumé visuel statique R^V est construit sous forme d'un ensemble d'images représentatives du contenu visuel de la vidéo. Cette représentation peut nous permettre d'avoir un accès direct aux différentes parties du document vidéo original. Elle peut également faire partie d'une interface interactive de recherche de données par le contenu. Les images composant le résumé seront alors considérées comme des index, ou des images requêtes pour les utilisateurs de cette interface. Cette collection d'images procure ainsi en un

clin d'œil une idée générale et globale des éléments pertinents compris dans cette vidéo. Cependant, cette représentation ne permet pas de capturer le dynamisme et la continuité des images d'une séquence vidéo.

- Un résumé visuel R^V peut aussi être présenté sous une forme dynamique. Cette représentation consiste à construire une séquence visuelle d'une durée désirée qui permet de préserver l'information temporelle des segments extraits de la vidéo. Ce type de résumé dynamique représente une version réduite du flux visuel de la vidéo entière. Autant ce genre de résumé préserve le dynamisme et l'évolution temporelle des données de la vidéo originale, autant il souffre du problème d'être linéaire. L'utilisateur est obligé de regarder la totalité du résumé d'une durée prédéfinie pour avoir une idée générale et comprendre le contenu de la vidéo originale.

La figure 4.1 illustre ces deux méthodes de représentation du résumé visuel R^V.

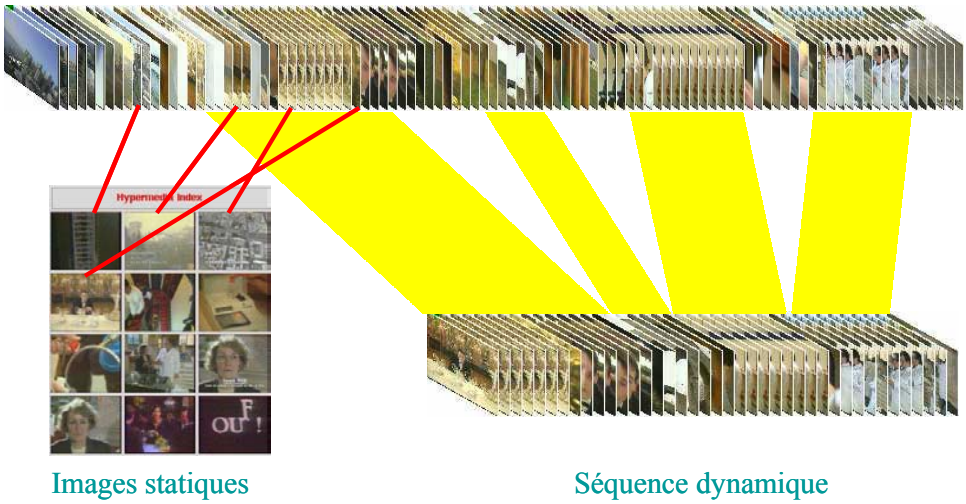

Figure 4.1: Représentation d'un résumé visuel R^V.

Dans l'intention de construire des résumés visuels, nous utilisons le principe de reconnaissance maximale PRM défini au chapitre précédent. Seules les informations visuelles extraites du flux vidéo considéré comme une succession d'images sont prises en considération dans ce chapitre.

L'application du PRM nécessite de préciser la nature de l'extrait employé, ainsi que la détermination de la règle de décision qui sera utilisée. Nous définissons un extrait E^V comme étant un segment de taille prédéfinie tiré aléatoirement du flux vidéo original. Cet extrait est une suite d'images ayant une relation de voisinage étroite.

La règle de décision permettra de valider si un extrait donné provient de la vidéo qui correspond au résumé créé ou non. Diverses règles de décision peuvent être envisagées. Nous avons décidé de définir une règle de décision raisonnable et automatisable. Cette règle est la suivante: «L'utilisateur décide qu'un extrait provient de la vidéo originale correspondant au résumé si au moins une image de l'extrait est similaire à une image appartenant au résumé».

Nous avons décidé d'adopter cette règle de décision car nous considérons que les images sont généralement distinctives. Il est très habituel que des téléspectateurs arrivent à identifier un film ou un programme à partir d'une seule image aperçue sur une affiche, dans une revue, un journal, ou à la télévision. Nous ne considérons qu'une vidéo unique à la fois sans utiliser d'autres vidéos. Tous les extraits E^V, utilisés lors de la sélection des images composantes du résumé ou lors de l'évaluation de ce dernier, sont tirés de la seule vidéo correspondante au résumé créé. Apercevoir une image de l'extrait similaire à une image du résumé est une preuve suffisante pour valider l'appartenance de cet extrait à la vidéo originale dont le résumé est montré.

Lors de notre étude, nous ne faisons pas de différence entre les résumés statiques et ceux dynamiques. La façon dont nous sélectionnons les éléments qui figureront dans le résumé final est indépendante du mode de présentation de ce dernier. Les résumés que nous présentons dans ce travail sont des résumés statiques composés d'un certain nombre d'images, ce nombre étant un paramètre désigné par l'utilisateur. A partir de ces résumés statiques, nous pouvons construire des résumés dynamiques en concaténant des segments de courtes durées, dont les centroïdes sont les images composant le résumé statique. La durée de ces segments dépend de la durée globale du résumé désirée par l'utilisateur ainsi que la durée minimale de chaque segment (clip). Si nous prenons des segments de très courte durée, le contenu du segment peut ne pas être compréhensible par l'utilisateur et le résumé ne serait pas agréable à voir. Pour cette raison il faut

fixer une durée minimale de six secondes par exemple pour chaque segment extrait de la vidéo originale. Afin d'éviter une transition brutale entre deux segments consécutifs, un effet de transition «fondu enchaîné» est utilisé pour adoucir le passage d'un segment à l'autre et lisser le résumé «videoskim».

4.3 Construction de résumé d'une vidéo unique

4.3.1 Expérience de reconnaissance Maximale

Ayant défini la règle de décision, nous adaptons notre principe de reconnaissance maximale au problème de la construction automatique de résumé d'une simple vidéo à la fois, en ne considérant que l'information visuelle. Pour ceci, nous proposons une expérience dont le scénario est le suivant {Figure 4.2}:

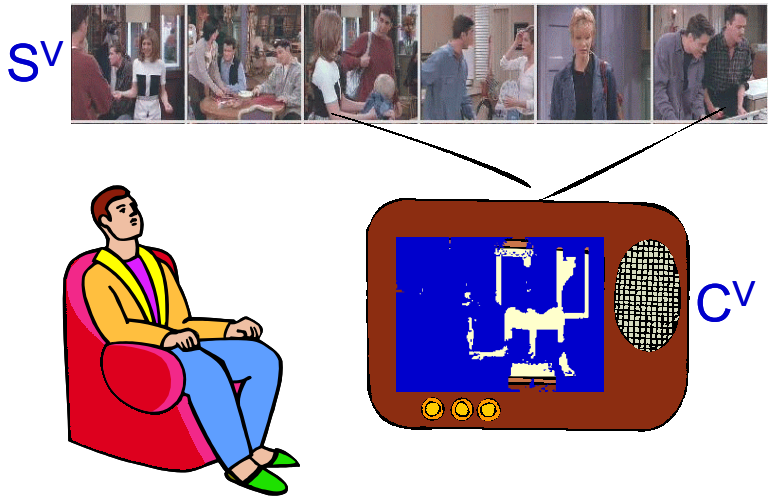

Figure 4.2: Scénario de l'expérience de reconnaissance visuelle maximale.

- Nous présentons le résumé vidéo R^V à l'utilisateur sous forme d'un ensemble d'imagettes.
- Ensuite, un extrait E^V, d'une durée prédéfinie d, choisi aléatoirement du flux vidéo est montré à ce dernier.
- Nous demandons à l'utilisateur de deviner si l'extrait E^V qu'il vient de voir provient de la même vidéo V qui correspond au résumé R^V ou non.
  Notre règle de décision correspond au comportement suivant de l'utilisateur:

- Si au moins une image de l'extrait E^V est similaire à une image du résumé R^V, il répond «oui» (et sa réponse est correcte).
- Si ce n'est pas le cas, il doute et ne peut se prononcer et fournir de réponse.

Le pourcentage de réponses correctes définit la performance de l'utilisateur dans cette expérience. Cette méthode est basée sur l'hypothèse que l'utilisateur a une mémoire visuelle parfaite des images du résumé. De plus, on suppose qu'il ne sait pas d'avance si les extraits montrés sont tirés de la même vidéo ou non (bien que ce soit le cas dans l'expérience).

Pour calculer la performance désirée d'une façon automatique, nous utilisons une définition de la similarité d'images, que nous essayons de définir de façon aussi cohérente que possible par rapport à l'appréciation des utilisateurs réels.

Notons que nous pouvons penser qu'une expérience complète devrait montrer à l'utilisateur des extraits pris de la vidéo originale ainsi que des extraits d'autres vidéos. Cependant, il est difficile de mettre ceci en pratique, parce que l'expérience serait dépendante du choix de ces autres vidéos. Si nous considérons la similarité des images avec une interprétation plutôt stricte, il est improbable qu'une image d'une première vidéo soit similaire à une image d'une autre vidéo (sauf dans le cas où le même clip est utilisé dans différentes vidéos); l'ajout d'autres vidéos ne changera donc vraisemblablement pas le nombre de réponses positives dans notre expérience.

Nous considérons que l'extrait est choisi d'une manière aléatoire dans la vidéo avec une distribution uniforme. Ceci donne la même importance aux différentes parties de la vidéo. On pourrait remettre ce choix en cause, par exemple en disant que les séquences du début de la vidéo sont plus importantes, ou que les séquences de la fin ne doivent pas être révélées. Il serait alors possible d'utiliser une distribution non-uniforme pour répondre à ces exigences.

Nous avons choisi de considérer que tous les extraits ont la même durée (qui est un paramètre dans notre processus de construction de résumés automatiques). Nous pourrions concevoir une expérience dans laquelle les deux paramètres que sont la position et la durée de l'extrait seraient choisis d'une manière aléatoire. Cependant, nous n'avons pas d'interprétation raisonnable concernant la variation de la durée, ni une distribution probabiliste judicieuse à suggérer.

4.3.2 Construction Automatique du résumé

Une foie définie cette expérience de reconnaissance maximale, nous avons besoin d'un processus de construction automatique de résumé vidéo qui assure une bonne performance (si possible optimale) pour cette expérience.

Faisant l'hypothèse que les extraits que nous considérons ont une durée d et que la vidéo comporte N images. Nous avons donc $N - d + 1$ différents extraits:

- E_1 contient les images $I_1, I_2, ...I_d$,
- E_2 contient les images $I_2, I_3, ...I_{d+1}$,
- et ainsi de suite jusqu'à E_{N-d+1} qui contient les images $I_{N-d+1}, I_{N-d+2}, I_N$.

Sachant que notre règle de décision est basée sur la similarité des éléments, nous faisons en premier lieu l'hypothèse que les images ont été classifiées en «classes de similarité», donc deux images sont considérées comme étant similaires si et seulement si elles appartiennent à la même classe:

$$I_i \; et \; I_j \; sont \; similaires \Longleftrightarrow C(I_i) = C(I_j) \tag{4.1}$$

C'est une hypothèse simplificatrice forte car le nombre choisi de classes ainsi que le seuil de ressemblance influeront d'une manière significative sur la qualité de la similarité visuelle des images appartenant à la même classe. Nous expliquerons ultérieurement dans la sous-section 4.3.4 comment ces classes de similarité sont construites.

La figure 4.3 illustre la relation entre les extraits, les images et les classes.

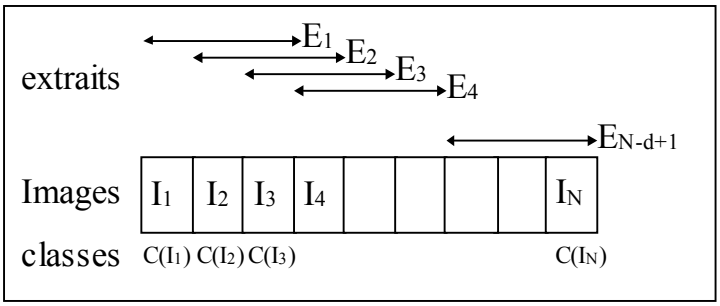

Figure 4.3: Relations entre extraits, images et classes.

Nous définissons la couverture $Cov(C)$ d'une classe C comme le nombre d'extraits qui contiennent au moins une image de la classe C:

$$Cov(C) = Card\{i : \exists j \quad I_j \in E_i \; et \; C(I_j) = C\} \tag{4.2}$$

De même, nous définissons la couverture $Cov(C_1, C_2, ...C_k)$ d'un ensemble de classes $C_1, C_2, ..., C_k$ comme étant le nombre d'extraits qui contiennent au moins une image appartenant à l'une de ces classes :

$$Cov(C_1, C_2, ...C_k) = Card\{i : \exists j, l \quad I_j \in E_i \; et \; C(I_j) = C_l\} \tag{4.3}$$

Avec ces notations, nous construisons notre résumé visuel R^V sous forme d'un ensemble d'images caractéristiques $I_1, I_2, ..., I_k$, où chaque image appartient à une classe donnée après la phase de classification. Nous définissons la performance Prf du résumé construit comme étant la couverture de l'ensemble des classes comportant les images du résumé créé divisée par le nombre de tous les extraits possibles:

$$Prf = \frac{1}{(N - d + 1)} Cov(C(I_1), C(I_2), ...C(I_k)) \tag{4.4}$$

Le résumé optimal R^V est celui qui est composé des images dont les classes maximisent la performance, c'est-à-dire:

$$R^V = \arg\max_{I_1, I_2, ..., I_k} \frac{1}{(N - d + 1)} Cov(C(I_1), C(I_2)...C(I_k)) \tag{4.5}$$

Par conséquent, la construction du résumé optimal par rapport à notre principe de reconnaissance maximale (PRM) peut être accomplie en deux étapes:

- En premier lieu, trouver un ensemble de classes avec une couverture maximale.
- Ensuite, sélectionner une image représentative de chaque classe.

4.3.3 Algorithme de Construction

Le résumé optimal de taille k peut être obtenu en faisant une énumération de tous les ensembles de k classes $\{C_1, C_2, ..., C_k\}$, puis en sélectionnant l'ensemble

ayant la meilleure couverture. Malheureusement l'énumération de la totalité des ensembles est en général trop coûteuse en temps de calcul: dans le cas d'un grand nombre de classes et d'une taille k assez importante, une explosion combinatoire est inévitable. En pratique, il est impossible de mettre en œuvre cette énumération si nous désirons construire un résumé vidéo en temps réel ou dans des temps raisonnables évitant aux utilisateurs des grands délais d'attente.

La création d'un résumé optimal par rapport à notre PRM n'est pas réalisable avec les moyens actuels, et nous allons par conséquent considérer la création d'un résumé sous-optimal. Nous présentons tout d'abord une heuristique qui permet la construction de résumés optimaux en faisant l'énumération d'un très grand nombre d'ensembles possibles sans énumérer la totalité, ce qui permet de diminuer le temps de calcul. Pour ceci, il est très judicieux de sélectionner minutieusement l'ordre dans lequel les classes sont sélectionnées, afin que la meilleure solution soit trouvée le plus rapidement possible.

Si une classe C_m est rajoutée à un ensemble existant $\{C_1, C_2, ..., C_{m-1}\}$, nous pouvons définir la couverture conditionnelle de la classe C_m par rapport à l'ensemble $\{C_1, C_2, ..., C_{m-1}\}$ comme étant la contribution de cette classe dans la couverture finale de cet ensemble :

$$\begin{aligned} Cov(C_m|C_1C_2...C_{m-1}) &= Cov(C_1C_2...C_m) - Cov(C_1C_2...C_{m-1}) \\ &= Card\left\{ \begin{array}{l} i : \exists j \ I_j \in E_i \ \ et \ \ C(I_j) = C_m \\ et \ \ \forall I \in E_i \forall j = 1, 2, ...m-1 \ \ C(I) \neq C_j \end{array} \right\} \end{aligned} \tag{4.6}$$

Si nous sélectionnons les classes dans l'ordre $C_1, C_2, ...C_k$, alors la couverture de l'ensemble des classes $\{C_1, C_2, ...C_k\}$ comportant les images qui composent le résumé sous-optimal peut être calculée en rajoutant à chaque étape de sélection la couverture conditionnelle de la classe insérée par rapport à l'ensemble courant des classes déjà sélectionnées:

$$Cov(C_1...C_k) = Cov(C_1) + Cov(C_2|C_1) + ... + Cov(C_k|C_1...C_{k-1}) \tag{4.7}$$

Maintenant que nous avons défini la couverture et la couverture conditionnelle d'une classe donnée, nous présentons l'algorithme que nous proposons pour la construction du résumé optimal.

Ce dernier procède comme suit:

Etape 1: Débuter le processus avec un ensemble vide de classes.

Etape 2: Ordonner les classes qui n'ont pas été encore sélectionnées en fonction de leur couverture conditionnelle décroissante par rapport à l'ensemble en cours.

Etape 3: Essayer d'ajouter à tour de rôle chaque classe à l'ensemble en cours.

Etape 4: Si la taille du résumé désirée est atteinte, remplacer la solution en cours par l'ensemble en cours s'il possède une meilleure couverture. Sinon revenir à l'étape 2 afin de continuer récursivement l'énumération.

Etape 5: Lorsque toutes les classes sont essayées, revenir à l'étape 3 de la récursion pour continuer l'énumération.

Durant la procédure de retour, il est possible d'éviter certaines énumérations et gagner un peu de temps de calcul en notant que la relation suivante est toujours maintenue si m<k :

$$Cov(C|C_1C_2...C_{m-1}C_m...C_k) \leq Cov(C|C_1C_2...C_{m-1}) \tag{4.8}$$

donc,

$$\begin{aligned} Cov(C_1C_2...C_{m-1}C_m...C_k) &= Cov(C_k|C_1C_2...C_{k-1}) + Cov(C_{k-1}|C_1C_2...C_{k-2}) + \\ &\quad ... + Cov(C_m|C_1C_2...C_{m-1}) + Cov(C_1C_2...C_{m-1}) \\ &\leq Cov(C_k|C_1C_2...C_{m-1}) + Cov(C_{k-1}|C_1C_2...C_{m-1}) + \\ &\quad ... + Cov(C_m|C_1C_2...C_{m-1}) + Cov(C_1C_2...C_{m-1}) \end{aligned} \tag{4.9}$$

Cette inéquation entraîne une borne supérieure pour la meilleure solution qui peut être construite en étendant l'ensemble $\{C_1, C_2, ...C_{m-1}\}$. Si la couverture avec cette borne supérieure est plus petite que la couverture de la meilleure solution courante, alors l'énumération peut être arrêtée à ce niveau. Ceci diminue le nombre de calculs requis, en préservant l'optimalité de l'algorithme.

Notons que l'algorithme commence par la sélection de la classe C_1 ayant la couverture maximale, puis C_2 qui a une couverture conditionnelle maximale par

rapport à la classe C_1 et ainsi de suite jusqu'à C_k. La première solution complète retrouvée est alors le résultat d'une série de choix avec le critère de type «greedy», c'est-à-dire la sélection de la meilleure solution en cours (un optimum local) à chaque étape d'insertion d'une nouvelle classe sans remettre en cause les choix précédents. Nous verrons par la suite qu'expérimentalement, c'est souvent le choix optimal parmi toutes les combinaisons possibles. Par conséquent, notre algorithme de construction de résumés vidéos sous-optimaux est en pratique basé sur la sélection d'images avec un critère de type «greedy».

Une fois que le meilleur ensemble de classes est calculé, il suffit de sélectionner une image représentative pour chaque classe. Sachant que les images de la même classe sont supposées similaires, le choix de l'image représentative n'influe pas sur la qualité du résumé. Dans notre étude, l'image représentative est sélectionnée comme étant celle dont le vecteur caractéristique est le plus proche du centre de la classe.

4.3.4 Classification des images

Notre méthode de construction de résumés vidéos est basée sur la similarité des images. Un extrait est jugé comme étant provenant de la vidéo originale correspondant au résumé présenté à l'utilisateur si et seulement si cet extrait comporte au moins une image similaire à une image du résumé.

Comme nous venons de le présenter dans les deux sous sections précédentes, nous avons décidé d'utiliser initialement la notion de classes de similarité comme un moyen de mesure et de décision de la similarité visuelle de l'ensemble des images. Dans ce paragraphe, nous présentons notre méthode de construction de ces classes.

Nous représentons les images considérées du flux vidéo original par des vecteurs caractéristiques. Nous utilisons deux types d'histogrammes pour capturer la distribution des couleurs de chaque image, ainsi que quelques informations locales. La première représentation consiste en des histogrammes de couleurs par région, et la deuxième en des histogrammes de blobs de couleurs [QvBS00]. De plus amples détails concernant la construction de ces vecteurs caractéristiques et les raisons du choix de l'une ou de l'autre caractérisation d'images seront donnés

dans la section consacrée à l'étude de la similarité des images. Une fois que les images sont représentées par des vecteurs caractéristiques, nous effectuons une classification de ces vecteurs caractéristiques sous forme d'un ensemble de classes d'images similaires. Chaque image n'appartient qu'à une seule classe à la fois, et est considérée comme similaire uniquement aux images de sa classe.

Cette classification stricte des images est réalisée de la manière suivante:

1. Commencer le processus de classification par un ensemble vide de classes.
2. Créer une première classe comportant une première image.
3. Calculer les centroïdes des classes actuelles.
4. Calculer les distances entre le vecteur caractéristique de la nouvelle image prise en compte et les centroïdes des classes existantes.
5. Si la distance minimale est supérieure au seuil de similarité défini, alors créer une nouvelle classe, sinon rajouter cette image à la classe dont le centroïde est le plus proche d'elle, revenir ensuite à 3.

Les images sont considérées dans un ordre aléatoire afin d'éviter un biais causé par l'ordonnancement temporel, car la construction d'une séquence vidéo fait que les images successives composant un plan sont souvent très similaires. Le seuil de similarité est choisi d'une façon expérimentale, plus d'explications seront présentées dans la section suivante. Une fois que cette étape initiale de classification est terminée, plusieurs étapes du type k-means sont réalisées afin de raffiner les classes. Nous considérons les images appartenant à la même classe sont visuellement similaires, cf. figure 4.4.

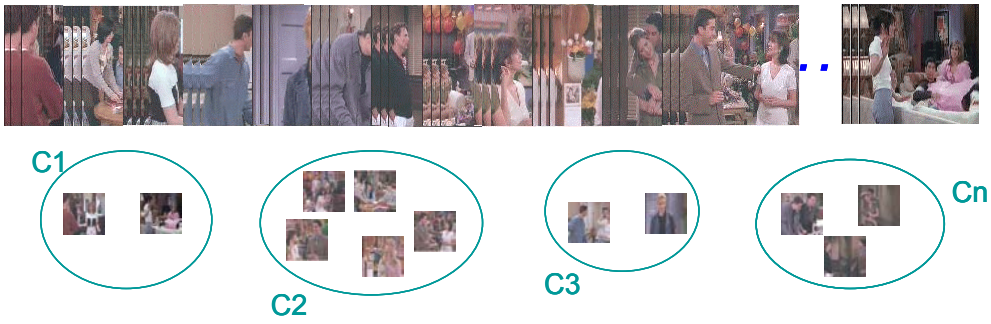

Figure 4.4: Classification stricte, avec un algorithme similaire à K-means.

4.3.5 Matrice de similarité des images

Nous avons été amenés à proposer une alternative à la notion de classes pour caractériser la similarité visuelle des images. Les raisons de ce choix sont détaillées ci-dessous.

Le choix du seuil de similarité (valeur en dessous de laquelle deux images sont considérées comme similaires) influe directement sur le nombre de classes obtenu lors de la phase initiale de la classification. Ceci implique que la qualité de la similarité visuelle au sein d'une même classe dépend d'une manière très significative du seuil de similarité choisi. Le choix du seuil est un problème très critique, et dépend directement du contenu visuel et des conditions d'enregistrement de la séquence vidéo traitée. Malheureusement, nous ne disposons pas d'un processus automatique de détermination du seuil le plus adéquat par rapport à chaque vidéo prise en considération. Si le seuil de similarité est petit, nous obtenons un grand nombre de classes de tailles réduites. Dans ce premier cas, le fait que les classes comportent un nombre réduit d'images garantit une certaine homogénéité et une bonne ressemblance du contenu visuel des images à l'intérieur de chaque classe. Par contre si le seuil est trop élevé, nous parvenons à un nombre limité de classes de tailles importantes. Dans ce cas, la similarité au sein d'une même classe perd de son sens.

Nous avons observé des effets de bord dans les classes construites, c'est-à-dire plusieurs images se situent au bord des classes. Les images de différentes classes se trouvant dans cette situation sont plus proches les unes des autres que du centre de leur classe. Ceci met en défaut notre hypothèse qui dit que les images appartenant à la même classe sont similaires entre elles et sont différentes des images appartenant aux autres classes.

Pour trouver une alternative à la classification, nous avons décidé de définir une nouvelle relation de similarité visuelle entre les images, basée uniquement sur la distance entre leurs vecteurs caractéristiques. Nous proposons pour cela de construire une matrice de similarité où chaque ligne correspond à une image donnée i. La ligne i comporte l'ensemble des images similaires à l'image i parmi toutes les images prises en considération, comme le montre la figure 4.5 .

Maintenant, que la notion de classes est remplacée par la notion de matrice de

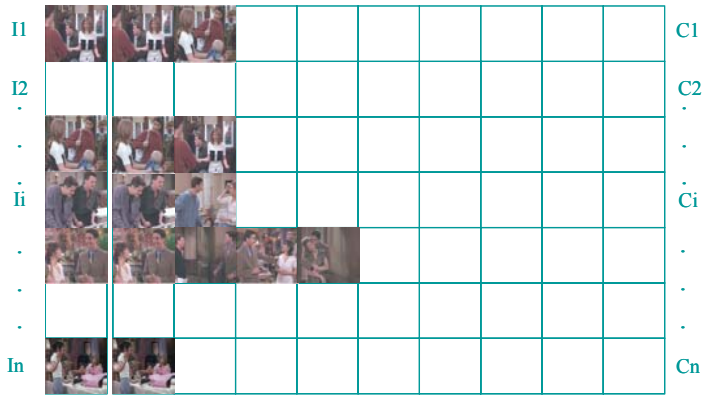

Figure 4.5: Matrice de similarité des images.

similarité, nous pouvons reprendre les formulations mathématiques en utilisant directement les images. Nous définissons la couverture d'une image I comme le nombre d'extraits comportant cette image ou une image qui lui est similaire, i.e. une image appartenant à la ligne de la matrice de similarité correspondante à l'image I.

$$Cov(I) = Card\{i : \exists j \quad I_j \in E_i \quad et \quad I_j \ \ similaire \ \ à \ I\} \tag{4.10}$$

De même, nous définissons la couverture d'un ensemble d'images $(I_1, I_2, ...I_k)$ comme étant le nombre d'extraits comportant au moins l'une de ces images, ou une image similaire à l'une d'elles.

$$Cov(I_1, I_2, ...I_k) = Card\{i : \exists j \ \ I_j \in E_i \ \ et \ \exists l, 1 \leq l \leq k \ \ et \ I_j \ similaire \ à \ I_l\} \tag{4.11}$$

Dans ce cas, le résumé optimal $\hat{R}^V$ est composé de l'ensemble de k images ayant la meilleure performance possible:

$$\hat{R}^V = \underset{I_1, I_2, ...I_k}{\arg\max} \frac{1}{(N-d+1)} Cov(I_1, I_2..., I_k) \tag{4.12}$$

Comme dans le cas où nous avons utilisé la classification, et afin de diminuer le temps de calcul, nous définissons la couverture conditionnelle d'une image I_m comme étant l'apport de cette image à un ensemble d'images présélectionnées $(I_1 I_2 ... I_{m-1})$

$$
\begin{aligned}
Cov(I_m|I_1I_2...I_{m-1}) &= Cov(I_1I_2...I_m) - Cov(I_1I_2...I_{m-1}) \qquad (4.13)\\
&= Card\left\{ \begin{array}{c} i : \exists j \ I_j \in E_i \ \ et \ \ I_j \ similaire \ à \ I_m \\ et \ \ \forall I \in E_i \forall j = 1, 2, ...m-1 \ \ I \ n'est \ pas \ similaire \ à \ I_j \end{array} \right\}
\end{aligned}
$$

Ceci implique que la couverture d'un ensemble d'images $(I_1...I_k)$ où les images sont sélectionnées dans l'ordre d'apparition est calculée de la manière suivante:

$$Cov(I_1...I_k) = Cov(I_1) + Cov(I_2|I_1) + ... + Cov(I_k|I_1...I_{k-1}) \qquad (4.14)$$

4.4 Etudes de la similarité des images

Comme nous l'avons vu tout au long des paragraphes précédents, notre méthode de création de résumés vidéos est basée sur le principe de reconnaissance maximale (PRM). Ce dernier dépend de la notion de similarité entre les éléments considérés pour la construction d'un résumé donné. Dans le cas de résumés visuels, il est lié directement à la notion de la similarité des images composant la vidéo traitée. Sachant que notre procédé de construction est automatique, nous avons besoin d'un processus de mesure et d'approximation de la similarité entre deux images assez proche du jugement des personnes sans aucune intervention humaine. A l'heure actuelle, la mesure de similarité visuelle d'une manière automatique constitue une question très critique dans le domaine du traitement d'images. Dans la littérature, plusieurs approches ont été proposées pour définir la similarité des images [SO96]. Il apparaît qu'il n'y a pas de méthode qui puisse réaliser ceci d'une manière idéale et parfaite vis-à-vis de la similarité visuelle jugée par un être humain. Cependant, chaque méthode présente quelques avantages, qui dépendent du contexte dans lequel l'étude de la ressemblance est requise. Par exemple, il y a une différence entre faire une comparaison d'images d'une même vidéo pour détecter les changements de plans, et identifier si deux images différentes contiennent le même objet ou la même personne. Dans ce travail, nous comparons une simple représentation d'histogrammes basée sur les régions, et une récente représentation utilisée dans le domaine de recherche d'images dans

les bases de données, appelée «histogrammes de blobs» [QvBS00] [How98]. Nous considérons deux représentations d'histogrammes afin de capturer la distribution de couleurs des images composant la vidéo; de même, deux mesures de distances sont étudiées dans le but de calculer la similarité de paires d'images extraites de la vidéo. Ces quatre éléments de comparaison sont détaillés ci-dessous.

4.4.1 Les histogrammes de couleurs

Les histogrammes de couleurs sont utilisés pour capturer la distribution des couleurs des pixels de chaque image. La similarité entre une paire d'images est calculée en comparant leurs histogrammes de couleurs respectifs. Ceci coïncide avec l'approche de Swain et Ballard [SB91] utilisée pour la recherche d'images dans des bases de données multimédias. Afin de capturer des informations locales, les images sont divisées en neuf régions égales, et un histogramme de couleur est calculé pour chaque région. Finalement, chaque image caractéristique est représentée par un vecteur construit en concaténant les neufs histogrammes de régions (voir figure 4.6). Pour le calcul de chaque histogramme de région, les valeurs des couleurs des pixels de la région représentées dans l'espace RGB sont quantifiées en 256 valeurs. Par conséquent, la taille du vecteur caractéristique représentant chaque image est égale à 256*9.

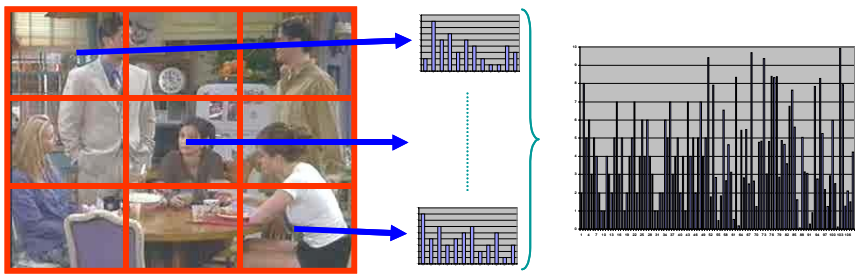

Figure 4.6: Construction d'un histogramme de couleurs par région.

4.4.2 Les histogrammes de blobs

Comme alternative à la représentation des histogrammes par région de couleurs, nous utilisons les histogrammes de blob. Qian et al.[QvBS00] ont proposé récemment une représentation d'histogrammes qui, au lieu de coder la distribution

fréquentielle de la couleur des pixels individuels, utilise un élément structurel afin d'inclure des informations locales dans l'histogramme.

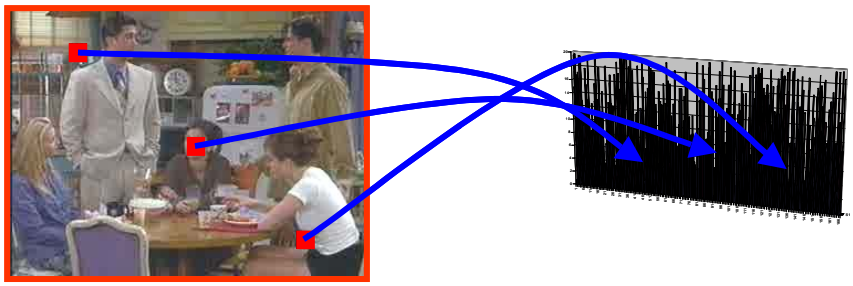

Figure 4.7: Construction d'un histogramme de blobs.

L'élément structurel, un carré k x k dans nos expériences, est déplacé à travers l'image et les groupes de pixels ayant une couleur uniforme dans cet élément sont appelés «blobs». Nous construisons des histogrammes de blobs en utilisant l'espace de couleur HSV. Cet espace est plus proche de la perception humaine des couleurs que les autres espaces de couleurs.

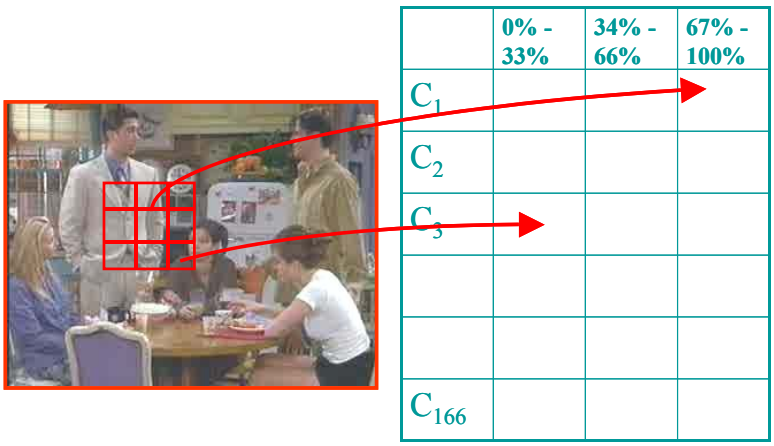

Figure 4.8: Représentation détaillée de la construction d'un histogramme de blobs de taille 166*3.

Les valeurs des couleurs des pixels de l'image sont quantifiées en 166 valeurs HSV et le pourcentage des pixels de chaque couleur dans un carré de taille nxn est quantifié en trois valeurs: {0-33% ; 34-66% ; 67-100%}, comme le présente la figure 4.8. La taille de l'histogramme des blobs est donc 166x3.

Il exsiste un grand nombre de distances pour comparer des histogrammes, nous avons opté pour les deux plus courantes:

La distance de Manhattan

La norme familière L1 peut être écrite comme suit :

$$L_1 = (P_D, P_M) = \sum_i |P_D(i) - P_M(i)| \tag{4.15}$$

où $P_D(i)$ et $P_M(i)$ sont les valeurs des histogrammes normalisés tel que chaque bin est divisé par le nombre total d'entrées dans l'histogramme.

La distance Euclidienne

De même, la norme L2 est définie comme suit:

$$L_2 = (P_D, P_M) = \sqrt{\sum_i (P_D(i) - P_M(i))^2} \tag{4.16}$$

4.4.3 Expériences sur la similarité visuelle

La meilleure façon d'évaluer la similarité visuelle entre deux images consisterait à demander l'avis des utilisateurs réels. Cependant, nous désirons ici construire des résumés vidéos de façon purement automatique (sans intervention humaine). Par conséquent, il faut définir une mesure mathématique capable de nous aider à porter un jugement «artificiel» sur la ressemblance des images. En revanche, nous avons besoin d'une distance et d'un seuil de décision; si la distance entre les vecteurs caractéristiques représentant les deux images est inférieure au seuil désigné, alors les deux images sont considérées comme similaires.

Nous mettons en œuvre avec l'intervention d'utilisateurs réels une expérience à travers laquelle, nous testons et comparons d'une part deux représentations du contenu visuel de chaque image, les histogrammes par région et les histogrammes de blobs et nous comparons d'autre part les deux types de distances, la distance Manhattan L1 et la distance Euclidienne L2.

L'objectif final de cette expérience est de trouver une valeur optimale du seuil de similarité pour les images des vidéos traitées en fonction des représentions et des distances étudiées.

- **Paramètres des expériences**

  Nos expériences se sont déroulées comme suit: nous avons calculé tout d'abord les vecteurs caractéristiques des images de la vidéo en utilisant des histogrammes par région ainsi que des histogrammes de blobs de différentes tailles. Nous avons testé des éléments structurels carrés ayant les dimensions suivantes: $3, 5, 7, 9, 11, 13, 15, 20, 30, 40, 50, et\ 100$. Ensuite, 200 paires d'images ont été sélectionnées à partir de plusieurs vidéos d'une manière aléatoire, avec la seule contrainte que les distances entre les histogrammes par région des deux images qui composent les paires sélectionnées soient distribuées d'une manière uniforme à travers un nombre de plages de distances L2 ($[0-100],]100-200]$, etc...). De manière analogue, 200 autres paires d'images ont été sélectionnées mais en utilisant la distance L1($[0-10],]10-20]$,etc...).

Figure 4.9: Interface d'évaluation de similarité visuelle.

- Pour l'ensemble des 400 paires d'images sélectionnées précédemment (200+ 200), un petit nombre d'utilisateurs (7 membres du labo) ont été invités à

déterminer si les deux images étaient visuellement similaires ou non. Ceci pour associer un jugement humain de similarité à l'ensemble des 400 paires d'images, comme le montre l'exemple de la figure 4.9. Une fois cette affectation effectuée, les différentes distances entre les paires d'images sélectionnées ont été calculées, en examinant toutes les combinaisons possibles des représentations (histogrammes par région et histogrammes de blobs) et des mesures des distances d'histogrammes (distance de Manhattan et distance Euclidienne).

On évalue le taux d'erreur de classification pour différentes valeurs du seuil de similarité. Ce taux d'erreur est défini comme le nombre d'images non similaires pour lesquelles la distance est inférieure au seuil, plus le nombre des paires similaires qui ont une distance plus grande que le seuil, par rapport à l'ensemble des paires prises en considération.

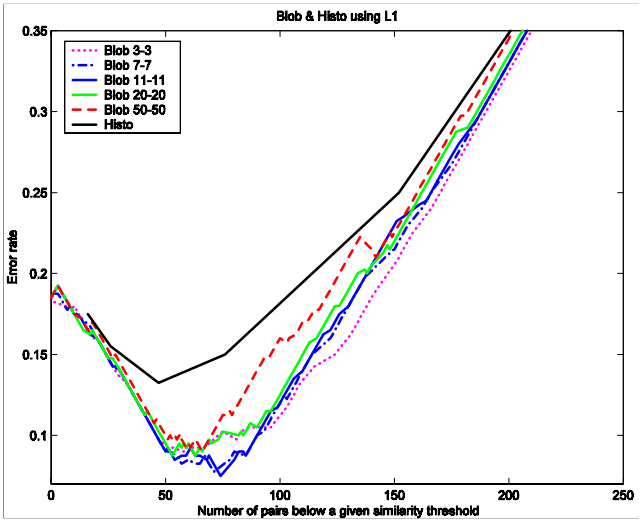

Figure 4.10: Taux d'erreur pour chaque plage de seuils.

- **Comparaison suivant le type d'histogramme**

  Le graphe de la figure 4.10 représente le taux d'erreur dans la tâche de classification d'images par rapport à leur similarité en fonction du nombre de paires d'images pour lesquelles la distance entre les deux images est inférieure à un seuil donné. Il est intéressant de souligner que la représentation des blobs provoque des taux d'erreur plus faibles que ceux obtenus

pour les histogrammes par région standards. De plus, la meilleure performance de la comparaison des 800 images des six vidéos (de tailles égales: 320x240) est obtenue pour les histogrammes de blobs construits avec un élément structurel carré (un blob) de taille 11x11 pixels.

- **Comparaison suivant le type de distance**

  La figure 4.11 représente les résultats de la comparaison pour les deux mesures de distances considérées. Dans le graphe, le taux minimum d'erreur obtenu avec les deux normes L1 et L2 est représenté pour des blobs de tailles 3x3 à 100x100. Les résultats des histogrammes par région sont présentés comme étant ceux du blob de taille 0x0 (0 sur l'axe horizontal) sur cette figure. Nous observons que pour les blobs ayant une taille inférieure ou égale à 40x40, les meilleurs résultats sont obtenus avec la norme L1. De plus, ces courbes suggèrent qu'il est conseillé de représenter les images de la vidéo en utilisant les histogrammes de blobs de couleurs calculés avec un élément structurel de taille égale à 11x11 ou 13x13 pixels. Les seuils qui correspondent au taux d'erreur minimal de 0.075 pour les histogrammes de blobs de tailles égales à 11x11 ou 13x13 sont respectivement 455 et 520. Parmi les 400 paires d'images, 74 paires ont une distance inférieure à 455 pour des blobs 11x11, et 76 ont une distance inférieure à 520 pour des blobs de 13x13. Dans le but de diminuer le temps de calcul lors de la construction des histogrammes de blobs, nous avons opté pour les blobs de taille plus petite parmi les deux meilleurs, c'est à dire les blobs 11x11 à la place de 13x13.

- **Une alternative pour le choix du seuil de similarité**

  Notre principe de reconnaissance maximale sur lequel est basé notre méthode de construction de résumé vidéo dépend particulièrement de la mesure de similarité des images. Dans le but de valider nos résultats à propos des différentes représentations des images et la comparaison des métriques, nous avons étudié davantage la qualité de la classification (similaire ou non) à travers l'ensemble des images consécutives de la vidéo. Il est évident que les images consécutives appartenant au même plan sont souvent très similaires. Ceci doit par conséquent nous aider à déterminer d'une autre façon le seuil

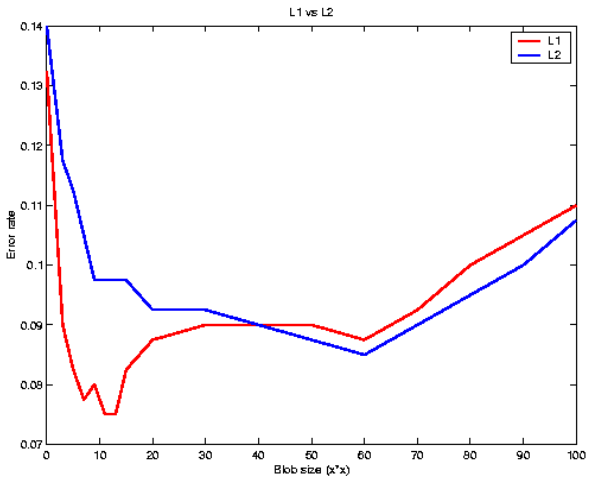

Figure 4.11: Taux d'erreur minimum en fonction de la taille de blob et la distance utilisée.

le plus approprié pour la similarité des images. La figure 4.12 présente l'histogramme des distances de tous les couples d'images consécutives pour trois vidéos de test. A partir de ces courbes, nous pouvons constater que la plupart des images consécutives ont une distance d'environ 200. Cependant, cette valeur est très stricte pour évaluer et juger la similarité d'une paire d'images. Un seuil d'une valeur de 300 ou 400 est probablement plus approprié. Afin de raffiner le choix du seuil, nous avons sélectionné aléatoirement les paires d'images pour lesquelles la distance de leurs histogrammes de blobs est proche de différentes valeurs potentielles du seuil. Ces paires d'images ont ensuite été évaluées par des utilisateurs réels afin de détecter, et d'éliminer les valeurs de seuil inappropriées. Finalement, nous avons choisi un seuil de similarité d'images de 455 pour la construction des résumés de vidéos.

En conclusion, ces études de la représentation et de la mesure de distance les plus adaptées pour la similarité des images nous ont conduit à utiliser les histogrammes de blobs de couleurs de la taille 11x11, en association avec la distance Manhattan. Pour confirmer que les blobs sont plus appropriés à notre méthode de construction que les histogrammes de régions, nous allons conserver dans un

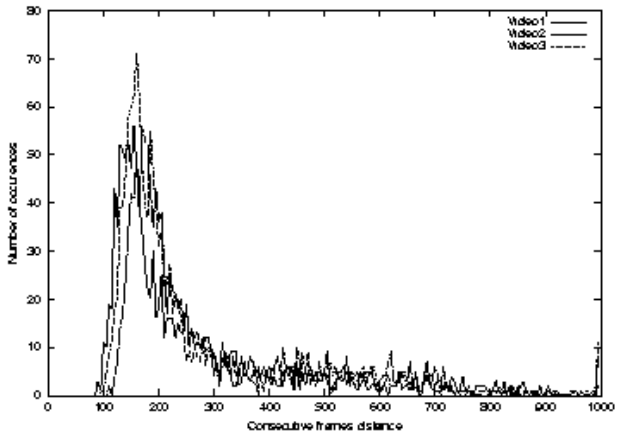

Figure 4.12: Histogramme des distances des couples d'images successives.

premier temps les deux représentations, puis comparer les performances des résumés des vidéos obtenus. Plus de détails sur cette question seront présentés dans la section 4.6.

4.5 Principe de l'Utilisateur Simulé

Jusqu'ici nous avons présenté notre PRM dans le cas de l'utilisation singulière de l'information visuelle, ainsi que notre méthode de construction automatique de résumés vidéos optimaux par rapport à la tâche particulière de reconnaissance d'une vidéo originale à partir d'un simple extrait de quelques secondes. Nous avons aussi présenté la solution quasi optimale issue du principe «greedy» que nous retenons en pratique afin d'avoir un mécanisme de création de résumés vidéo dans des temps réels adaptés aux besoin des utilisateurs. Une fois que ces résumés sont construits, nous avons besoin d'un outil d'évaluation de leur qualité. Avant de présenter notre approche pour évaluer les performances des résumés obtenus après la phase de construction basée sur le PRM, faisons tout d'abord un petit rappel des diverses approches d'évaluation disponibles dans la littérature.

- Une première façon de procéder est de demander à un groupe d'utilisateurs d'évaluer directement la qualité des résumés ou d'effectuer une comparaison de plusieurs résumés créés par des méthodes distinctes. Un autre moyen

d'établir une évaluation réaliste consiste à demander à un groupe d'usagers de réaliser d'abord un certain nombre de tâches (par exemple répondre à des questions, ordonner quelques séquences, etc...) ; ceci peut se faire en ayant connaissance ou non du résumé correspondant à la vidéo traitée. Ensuite, on mesure l'impact de ce résumé sur les performances des utilisateurs lors de l' accomplissement de ces différentes tâches.

- Une deuxième façon consiste à utiliser un critère purement mathématique. La valeur calculée pour chacun des résumés est ensuite utilisée comme une mesure de la qualité de ce dernier.

D'une manière générale, les approches qui entraînent la participation de vrais utilisateurs sont celles qui donnent les résultats les plus réalistes, parce que nous pouvons avoir une idée claire des raisons qui expliquent la bonne ou la mauvaise qualité du résumé. Malheureusement, ces approches sont très coûteuses et très difficiles à mettre en oeuvre. Afin d'obtenir des résultats ayant une certaine signification statistique, plusieurs utilisateurs sont requis, et quand différents résumés doivent être comparés, un utilisateur qui a vu le résumé A ne peut évaluer un deuxième résumé B parce qu'il a déjà assimilé des informations qui vont l'influencer. Ceci restreint fortement le nombre d'expériences qui peuvent être effectuées avec un groupe d'utilisateurs, et il devient difficile de comparer plusieurs sous-variantes d'une même méthode.

Par conséquent, l'évaluation de la qualité du résumé est souvent réalisée à l'aide d'un critère mathématique. Dans ce cas, l'évaluation n'est ni coûteuse, ni ambiguë; elle permet des comparaisons statistiques, et peut être utilisée aussi souvent que souhaité. Une difficulté réside néanmoins dans le fait qu'il n'est pas toujours facile de comprendre le sens de cette mesure de performance, parce qu'elle n'est pas associée à une activité ou un comportement des utilisateurs réels. Par conséquent, la vraie importance d'une performance croissante reste mystérieuse pour les observateurs. Si nous souhaitons juger d'une manière réaliste la performance d'un résumé créé par notre méthode basée sur le PRM avec l'aide d'utilisateurs réels, nous devons présenter à chaque utilisateur l'ensemble des extraits possibles d'une durée prédéfinie, et lui demander de deviner si l'extrait qui lui est montré provient ou non de la vidéo qui correspond au résumé qu'il a sous les yeux. Cette évaluation réaliste est presque impossible en pratique, et nous

proposons donc d'utiliser un utilisateur «simulé», qui reproduit en quelque sorte le comportement des vrais usagers. Pour ceci, nous supposons qu'un utilisateur réel décide de l'origine d'un extrait en comparant uniquement son contenu visuel avec celui du résumé, et nous faisons abstraction des informations sémantiques ou intuitives que le vrai utilisateur utilise réellement pour établir son jugement. De cette manière, nous combinons les avantages des deux types d'évaluation. Nous évaluons nos résumés en utilisant une méthode basée sur le même principe de reconnaissance maximale que la méthode de construction. Notre méthode d'évaluation se base sur un critère mathématique qui nous permet de simuler le comportement des utilisateurs réels sans avoir recours à ces derniers. Nous avons proposé une tâche réaliste aux utilisateurs, pour laquelle une mesure de performance et une règle de décision ont été établies. Grâce à certaines hypothèses réalistes, il a été possible de prédire le comportement des utilisateurs lors de la réalisation de cette tâche. Nous allons voir ainsi que ce modèle de comportement simulé peut être utilisé efficacement et donc utilisé par notre algorithme de construction et d'évaluation automatique de résumés vidéos.

En résumé, la méthode d'évaluation que nous proposons, basée sur l'utilisateur simulé consiste à calculer la performance du résumé construit de la manière suivante. Pour chaque extrait de durée prédéterminée tiré d'une vidéo, nous comparons toutes les images qui le composent avec toutes les images qui constituent le résumé en utilisant une mesure mathématique. Si une image de l'extrait au moins est similaire à une image du résumé (si leurs vecteurs caractéristiques sont proches), nous incrémentons le nombre d'extrait reconnus à l'aide du résumé. Une fois que tous les extraits possibles sont traités, nous divisons le nombre d'extraits reconnus par le nombre d'extraits possibles pour obtenir la performance du résumé. La quantité mathématique obtenue représente le taux de reconnaissance d'une vidéo originale en ayant uniquement connaissance de son résumé, par le moyen d'un extrait tiré de la vidéo originale d'une manière aléatoire.

4.6 Expériences

Dans le but de mettre en œuvre notre méthode de construction des résumés vidéos basée sur le PRM en utilisant uniquement l'information visuelle, nous

avons effectué plusieurs expériences. Au cours de ces dernières, nous avons évalué la qualité des résumés résultant de notre processus de création à l'aide de notre utilisateur simulé, lequel était conçu à partir d'hypothèses sur le comportement des utilisateurs réels. Afin de réaliser ces expériences, nous avons utilisé un ensemble de vidéos de type $Mpeg1$. Les vidéos prises en considération étaient:

- six épisodes de la série télévisée «Friends» (vidéos dénotées par F1, F2, F3, F4, F5 et F6),
- un documentaire intitulé «Histoire d'eau» (vidéo H),
- une épisode de fiction «Chapeau melon et bottes de cuir» (vidéo C).

Les épisodes de «Friends» ont été enregistrés dans notre laboratoire à partir de la chaîne de télévision France2. Les deux autres vidéos font partie d'un corpus de vidéos distribué par l'INA (Institut National de l'Audio-Visuel) dans le cadre de l'action indexation multimédia réalisée par le groupe de travail GT10 du GDR-PRC ISIS.

Sachant que nous sommes plus intéressés par une analyse globale des vidéos, que par une analyse détaillée sur une courte durée, nous avons sous-échantillonné les vidéos en retenant une image par seconde. Ensuite, nous avons construit pour chaque image retenue un vecteur caractéristique pour la représenter. Ce vecteur est construit sous la forme d'un histogramme par région ou d'un histogramme de blobs de couleurs 11x11. Dans le but de simplifier davantage le traitement et de diminuer les temps de calcul, les images consécutives dont les vecteurs caractéristiques étaient très proches (ayant une distance inférieure à un seuil très réduit) étaient confondues, et seulement la première image était conservée, aussi que le nombre des images consécutives confondues.

Comme nous l'avons mentionné dans la chapitre précédent, notre principe de construction est basé sur la notion de similarité des éléments composant le document original, dans ce cas la similarité des images. Nous avons adopté deux procédés distincts pour caractériser la similitude entre les images composant le flux vidéo. Ces deux procédés reposent sur le calcul des distances entre les vecteurs caractéristiques représentant les différentes images. Pour le premier, l'ensemble des images gardées après les phases préliminaires de sous-échantillonnage et de fusion ont été classifiées sous forme de classes de similarité;

pour le second procédé, une matrice de similarité a été construite.

Nous avons examiné les résultats de notre processus de création automatique de résumés vidéos en combinant les deux procédés de définition de la similarité avec les deux représentations d'images proposées.

Dans la suite de cette section, nous développons tout d'abord les expériences réalisées en utilisant les histogrammes de couleurs par région comme représentation caractéristique de chacune des images considérées. Ensuite, nous exposons les résultats des expériences effectuées avec la deuxième représentation choisie (histogrammes des blobs). Dans les deux cas, nous considérons les deux types de caractérisation de la similarité (soit des classes de similarité obtenues par un processus de classification, soit une matrice de similarité).

4.6.1 Histogrammes par région

Nous avons vu, dans l'étude de la similarité visuelle des images présentée dans ce même chapitre, comment nous pouvons choisir la mesure ainsi que le seuil de similarité les plus adaptés à nos expériences. Il s'avère que lorsque les images sont représentées par des histogrammes de couleurs par région, c'est la distance L1 combinée avec un seuil de similarité égal à 100 qui permet de produire les décisions de similarité les plus similaires au jugement humain.

4.6.1.1 Avec classification

Une fois que les images sont représentées par des histogrammes de couleurs par régions, les vecteurs caractéristiques construits pour chaque vidéo sont classifiés avec la méthode exposée dans la section 4.3.4. Le tableau 4.1 ci-dessous présente la durée en secondes des différentes vidéos ainsi que le nombre de classes de similarité, résultant de la phase de classification.

	F1	F2	F3	F4	F5	F6	H	C
Durée en secondes	1290	1331	1310	1298	1464	1394	2118	3035
Nombre de classes	252	292	248	165	226	286	933	1065

Tableau 4.1: Durée des vidéos et le nombre de classes respectives

Le résumé de chaque vidéo est construit indépendamment des autres vidéos. La construction de chaque résumé est réalisée suivant l'algorithme que nous avons présenté dans la sous-section 4.3.3. Une première solution est obtenue à l'aide d'un processus de type «greedy». Afin d'évaluer la qualité du résumé ainsi obtenu, par rapport à la solution optimale, nous remettons cette solution en cause. La méthodologie employée est inspirée de l'algorithme «branch and bound». Chaque élément du résumé (du dernier au premier élément) est substitué par l'élément ayant le rang suivant en terme de couverture conditionnelle.

Les temps de construction de ces résumés avec «greedy» uniquement sont généralement proches d'une minute, pour des vidéos d'environs 20 minutes. Lorsque le calcul du résumé optimal est mis en œuvre, les temps de construction augmentent considérablement. Il faut dans ce cas plusieurs heures, voir même plusieurs jours pour réaliser le résumé (ceci varie en fonction de la durée de la vidéo et la taille des extraits utilisés lors du calcul des couvertures).

Pour chacune des vidéos, nous avons construit plusieurs résumés en utilisant les tailles d'extraits suivantes: $4, 6, 8, 10, 20$ et 40 secondes. Lors de nos différentes expériences, les résumés obtenus avec l'algorithme «branch and bound» n'ont pas amélioré la couverture des résumés, à l'exception d'une seule vidéo parmi les huit utilisées. Pour cette vidéo, en utilisant des extraits de taille égale à 40 secondes, une meilleure solution était enregistrée après l'utilisation du «branch and bound». Ceci indique que les résumés obtenus sont les plus efficaces par rapport à la tâche que nous avons définie, laquelle reproduit relativement bien notre principe de reconnaissance maximale. En d'autres termes, les résumés obtenus sont ceux qui permettent à notre utilisateur simulé de reconnaître le maximum d'extraits tirés d'une vidéo donnée, en ayant connaissance seulement de son résumé. Suite à ces observations, il apparaît qu'en pratique, une énumération de l'ensemble de toutes les solutions possibles n'est pas indispensable pour avoir des résumés quasi-optimaux.

La figure 4.13 présente la couverture des résumés (exprimée en pourcentage) pour des extraits de différentes durées.

- En analysant ce graphe, nous remarquons comme nous l'avons prévu que la couverture de chaque résumé augmente en fonction de la durée des extraits. Ceci nous semble logique puisque avec l'augmentation de la durée

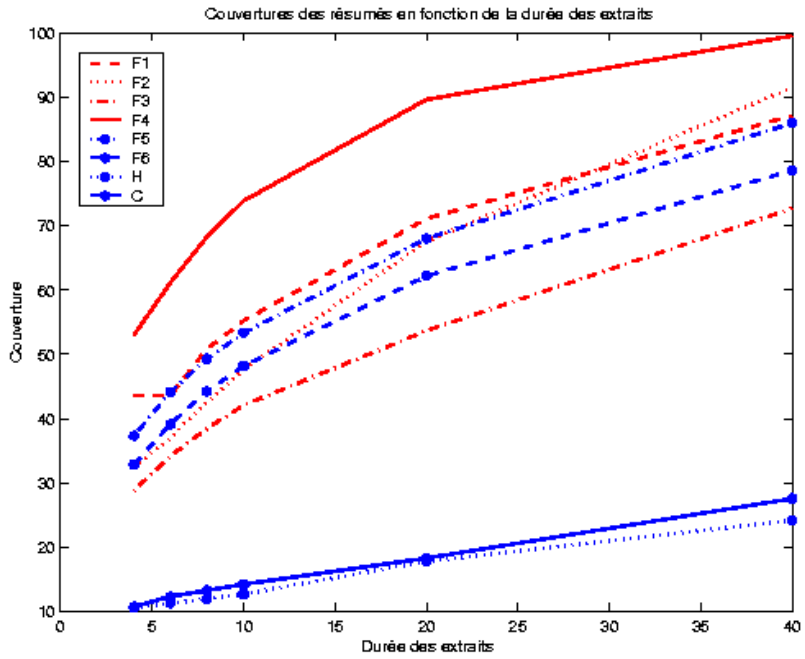

Figure 4.13: Couvertures des résumés de taille 6 des vidéos considérées en fonction de la durée des extraits (histogrammes par région et classification).

de l'extrait utilisé, la probabilité qu'un extrait tiré aléatoirement de la vidéo contienne au moins une image similaire à une image du résumé augmente. Nous observons aussi que la couverture obtenue diffère d'une vidéo à l'autre. Les résumés des épisodes de la série télévisée donnent des couvertures élevées, débutant entre 28% et 53%, alors que le documentaire et l'épisode de fiction donnent des couvertures qui démarrent seulement à 10%. Ceci est dû principalement au fait que le documentaire et l'épisode de fiction comportent une plus grande diversité (un grand nombre de classes de similarité) que les épisodes de la série télévisée. Il faut noter que cette divergence de nombre de classes est aussi provoquée par le seuil de similarité utilisé. Ce dernier avait été calculé lors de nos expériences d'étude de similarité, où uniquement les épisodes de la série télévisée «Friends» étaient prises en considération. En ce qui concerne ces épisodes, nous pouvons distinguer la quatrième vidéo (F4) par rapport aux autres. Le résumé correspondant à cette vidéo enregistre les meilleures performances. Notons

que pour les cinq autres épisodes, les nombres des classes obtenues pour chacune des vidéos après la phase de classification se regroupent, et sont aux alentours de 250. Cependant, les images de la vidéo F4 sont classifiées en seulement 165 classes. Nous expliquons le phénomène d'avoir des performances faibles lorsque le nombre des classes est élevé (et vice versa) par le fait que la probabilité qu'un extrait comporte une image de la même classe qu'une autre image du résumé est plus importante chaque fois que le nombre de classes est réduit.

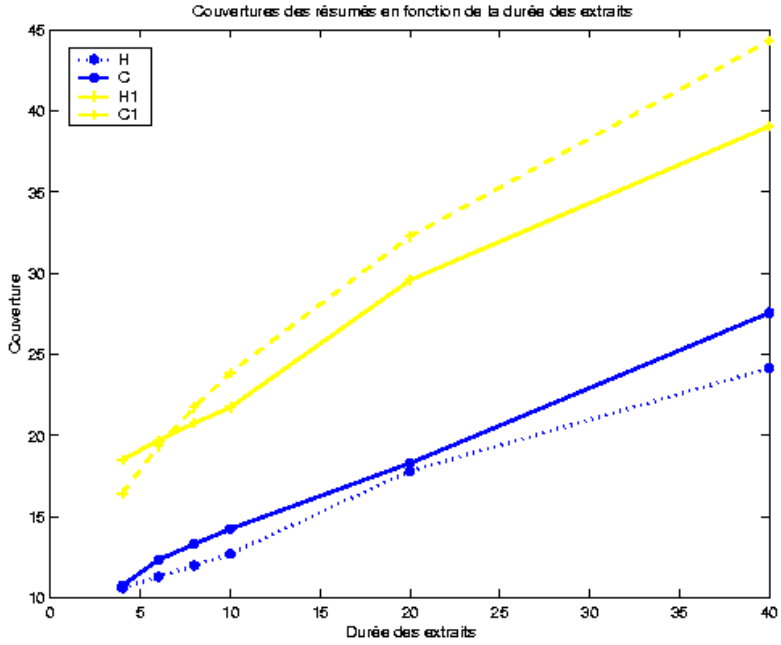

Figure 4.14: Couverture des résumés de taille 6 en fonction de la durée des extraits pour H,C et leurs versions tronquées H1,C1 (histogrammes par région et classification).

- Un autre facteur à prendre en compte est que les durées des épisodes de la série télévisée «Friends» sont assez courtes par rapport aux deux autres vidéos. Afin de déterminer l'influence de la durée sur les performances de notre résumé, nous avons considéré deux nouvelles vidéos H1, C1. Ces deux vidéos sont des sous-séquences vidéos obtenues en tronquant les vidéos H, C après une durée comparable à la durée des six autres vidéos.
  Le tableau 4.2 indique les durées ainsi que les nombres de classes des deux

	H	C	H1	C1
Durée en secondes	2118	3035	1284	1287
Nombre de classes	933	1065	541	605

Tableau 4.2: Durée des vidéos et le nombre de classes respectives pour H, C et H1, C1

nouvelles vidéos tirées du documentaire et de la série de fiction. D'après ce tableau, nous notons que le nombre de classes a été presque réduit de moitié pour les vidéos H1 et C1. La figure 4.14 présente les performances des résumés construits pour ces vidéos en fonction de la taille des extraits. Nous remarquons que les performances des résumés des vidéos les plus courtes sont meilleures que celles des vidéos longues. Nous pensons que cette amélioration est due principalement au fait que le nombre de classes a considérablement diminué.

4.6.1.2 Avec matrice de similarité

Dans cette section, nous présentons les résultats de nos expériences dans le cas où la similarité des images est caractérisée sans recourir à une classification stricte. Nous déterminons pour chaque image l'ensemble des images qui lui sont similaires en calculant les distances entre le vecteur caractéristique d'une images donnée et tous les vecteurs correspondant au reste des images. Les distances mesurées sont ensuite comparées à un seuil de similarité. Deux images sont considérées comme similaires si la distance entre leurs vecteurs caractéristiques est inférieure au seuil prédéfini. Une matrice de similarité est construite de la façon présentée dans la section 4.3.5.

Lors des expériences précédentes, où une classification stricte était utilisée pour définir la similarité des images, nous avons constaté que la première solution retrouvée par le processus de type «greedy» est souvent la plus optimale. Pour diminuer les temps de calcul, nous avons par conséquent décidé, lors des expériences suivantes d'arrêter le processus de construction dès que la première solution a été générée par le critère «greedy». Donc, à chaque étape de la construction du résumé vidéo, l'image ayant la meilleure couverture conditionnelle

par rapport au résumé en cours est sélectionnée, et ainsi de suite jusqu'à ce que la taille désirée du résumé (en nombre d'images) est atteinte.

Nous avons pour ces expériences utilisé les six épisodes de la série «Friends», ainsi que les deux sous-séquences vidéos extraites du documentaire «histoire d'eau» et de la série de fiction «Chapeau melon et botte de cuir» de durées équivalentes aux six premières vidéos. La matrice de similarité était construite en utilisant la distance L1, et la même valeur de seuil de similarité (100). Les résumés que nous avons ainsi construits sont composés chacun de six images. La taille du résumé est un paramètre dans notre processus de construction, qui varie en fonction du besoin des utilisateurs. Le graphe de la figure 4.15 représente les couvertures des résumés obtenus en fonction de la durée des extraits utilisés lors de la construction. Ces résumés sont construits à partir des images caractérisées par des histogrammes de région dont la relation de similarité est définie par une matrice de similarité.

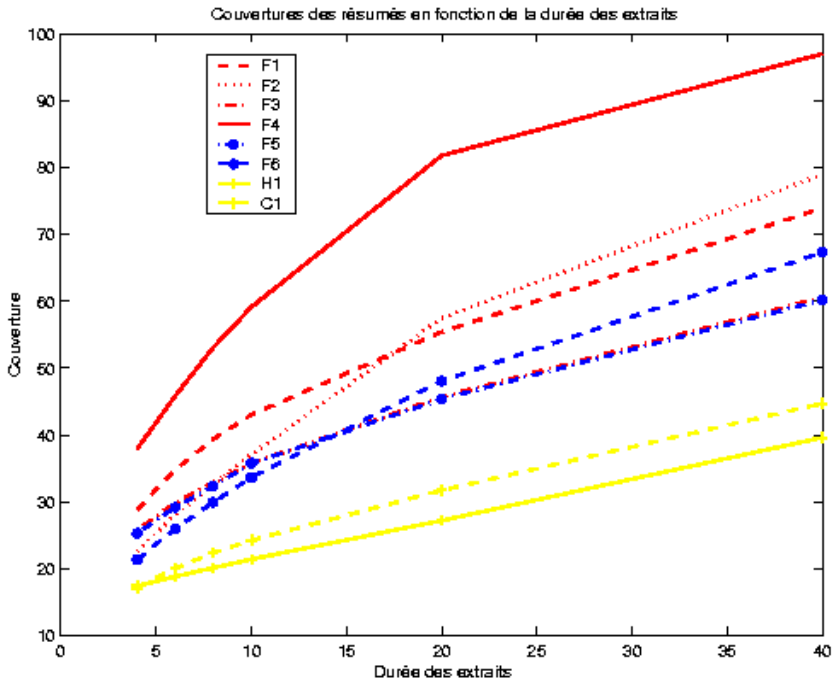

Figure 4.15: Couvertures des résumés de taille 6 des vidéos considérées en fonction de la durée des extraits (histogrammes par région et matrice de similarité).

De même que dans le cas de la classification, nous observons que la couverture

de chaque résumé augmente en fonction de la durée des extraits. Ceci est dû à la même raison que dans le cas précédent, c'est à dire qu'avec l'augmentation de la durée de l'extrait utilisé, la probabilité qu'un extrait tiré aléatoirement de la vidéo contienne au moins une image similaire à une image du résumé agmente. D'une manière analogue au cas précédent, la couverture des vidéos diffère de l'une à l'autre. Les résumés des épisodes de la série télévisée donnent cette fois-ci des valeurs de couvertures plus faibles que celles de la figure 4.13, débutant entre 20% et 40%. Les parties tirées du documentaire (H1) et de l'épisode de fiction (C1) donnent des couvertures comparables à celles des épisodes (F1 à F6) mais qui restent entre 18% et 40%. Les performances des résumés dépendent directement de l'homogénéité de leur contenu visuel. La divergence des performances des résumés des différentes vidéos s'accentue avec la taille des extraits utilisés. Nous expliquons ceci de la manière suivante: pour une vidéo dont le contenu est homogène, chaque fois que la taille de l'extrait augmente, la probabilité que cet extrait contienne une image similaire à une image insérée dans le résumé augmente plus rapidement que pour une vidéo dont le contenu visuel est très hétérogène.

4.6.2 Histogrammes des blobs

Dans le cas des histogrammes de blobs, les expériences réalisées pour étudier la similarité visuelle des images ont montré que la meilleure approximation de cette similarité est obtenue avec la distance L1 et un seuil de similarité égal à 455.

4.6.2.1 Avec classification

Le tableau 4.3 ci-dessous présente la durée en secondes des différentes vidéos, ainsi que le nombre de classes de similarité après une classification stricte des images de chaque vidéo, en utilisant un seuil de similarité approprié (455).

	F1	F2	F3	F4	F5	F6	H	C
Durée en secondes	1290	1331	1310	1298	1464	1394	2118	3035
Nombre de classes	76	75	97	60	117	83	1003	1070

Tableau 4.3: Durée des vidéos et le nombre de classes respectives

De même que dans le cas des histogrammes par région, nous construisons des résumés correspondant à l'ensemble des vidéos traitées. En premier lieu, nous considérons le cas où la similarité des images est caractérisée par une classification stricte. La construction de chaque résumé est réalisée suivant notre PRM. Nous prenons en compte la première solution sous-optimale produite par un processus de type «greedy ». Une fois que les classes ayant les plus grandes couvertures conditionnelles sont sélectionnées, chacune d'elle est représentée par l'image la plus proche de son centroïde. Ces images composent le résumé résultant de notre approche de construction. Lors du calcul de la couverture conditionnelle, nous avons testé plusieurs tailles d'extraits: $4, 6, 8, 10, 20$ et 40 secondes.

Le graphe de la figure 4.16 présente les couvertures (exprimées en pourcentage) des résumés construits en fonction des tailles des extraits utilisés.

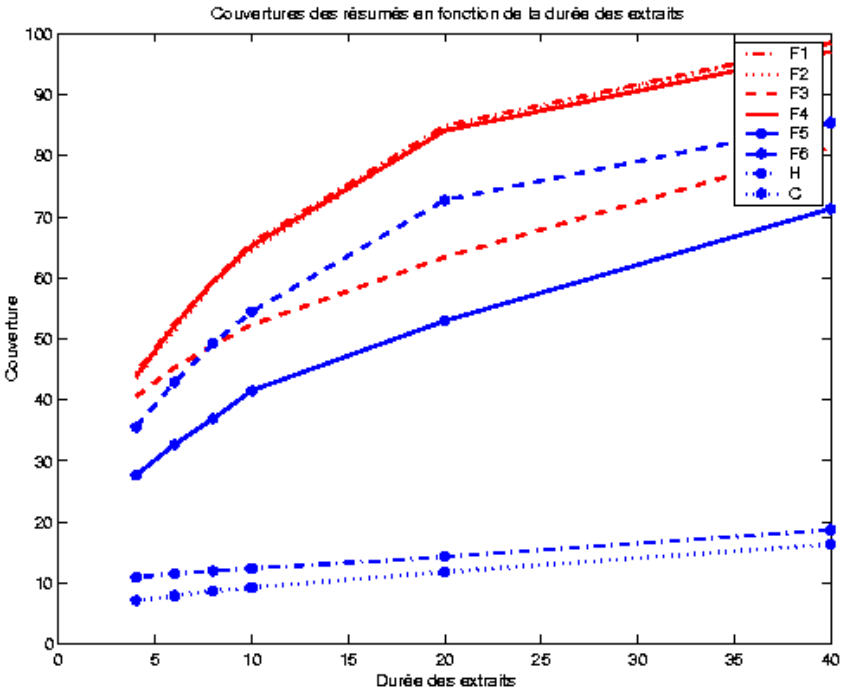

Figure 4.16: Couvertures des résumés de taille 6 des vidéos considérées en fonction de la durée des extraits (histogrammes de blobs avec classification).

La nouvelle représentation des images par les histogrammes des blobs n'a pas affecté la croissance de la couverture en fonction des tailles des extraits par rapport au cas des histogrammes par région. Chaque fois que la taille des extraits

augmente, la probabilité qu'un extrait comporte une image de la même classe qu'une autre image appartenant au résumé augmente. A part les deux vidéos F1 et F4 dont les tracés se superposent, nous observons que les couvertures obtenues diffèrent sensiblement d'une vidéo à l'autre. Les résumés des épisodes de la série télévisée donnent des couvertures élevées, débutant à peu près à 40%, alors que le documentaire (H) et l'épisode de fiction (C) donnent des couvertures qui démarrent seulement à 10%. Comme le montre le tableau 4.3, le nombre de classes construites pour les deux vidéos H et C est dix fois plus grand que le nombre de classes calculés pour les autres vidéos. Le fait d'avoir un très grand nombre de classes $(1003, 1070)$ diminue vraisemblablement la probabilité qu'un extrait donné comporte une des classes sélectionnées pour faire partie du résumé (d'autant plus que le nombre de classes sélectionnées dans ces expériences est égale à six seulement). Il faut noter que le nombre de classes n'est pas le seul paramètre qui influence le calcul des couvertures conditionnelles; la distribution des images tout au long de ces classes est aussi un facteur important. Si nous avons parmi ce grand nombre de classes, quelques classes de taille très importante, puis uniquement des singletons ou des classes de taille très réduite, la couverture globale sera importante. Dans ce cas, la plupart des images appartiennent à un nombre limité de classes, donc un grand nombre d'extraits comportera au moins une image de ces classes. L'insertion de quelques classes de taille importante dans le résumé global par notre algorithme de construction permettra d'obtenir des couvertures importantes, malgré un nombre total de classes important.

4.6.2.2 Avec matrice de similarité

Comme dans le cas précédent, nous gardons la première solution générée par le critère de type «greedy». A chaque étape de la construction du résumé vidéo, l'image ayant la plus grande couverture conditionnelle par rapport au résumé actuel est sélectionné et insérée dans le résumé, jusqu'à ce que nous atteignions la taille prédéfinie par l'utilisateur.

Nous avons utilisé les mêmes vidéos que dans le cas avec classification, ainsi que la même valeur du seuil de similarité 455 pour créer la matrice de similarité. Chaque ligne de cette matrice comporte toutes les images visuellement similaires (selon notre mesure) à l'image correspondant à cette ligne. Les résumés construits

sont composés de six images. Cette taille reste un paramètre dans notre processus de construction, et peut varier en fonction des besoins des utilisateurs.

Le graphe de la figure 4.17 présente les couvertures de résumés obtenues en fonction de la durée de l'extrait utilisé lors de la construction.

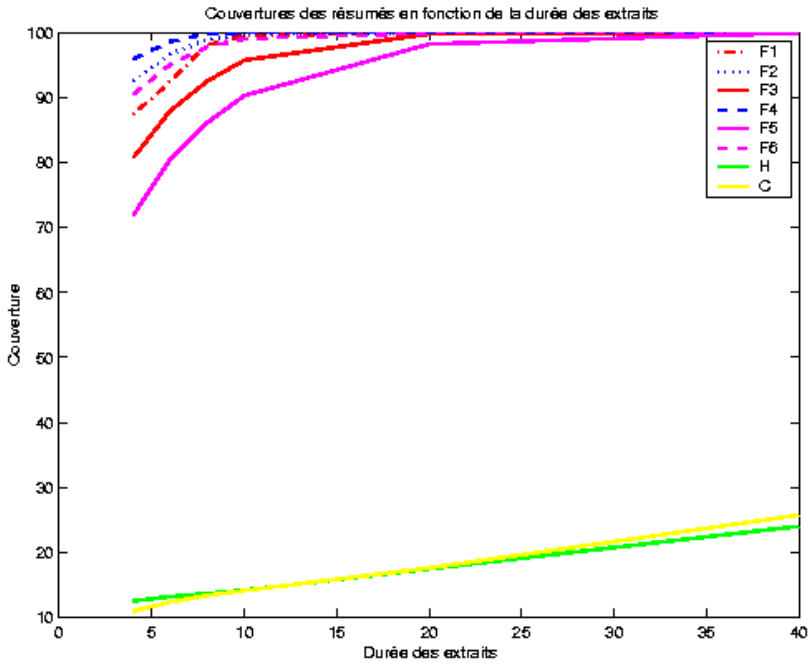

Figure 4.17: Couvertures des résumés de taille 6 des vidéos considérées en fonction de la durée de l'extrait (histogrammes de blobs et matrice de similarité).

Pour la même raison que dans le cas de la classification stricte, nous observons que la couverture de chaque résumé augmente en fonction de la durée des extraits. C'est-à-dire avec l'augmentation de la durée de l'extrait utilisé, la probabilité qu'un extrait pris aléatoirement de la vidéo contienne au moins une image similaire à une autre parmi celles qui composent le résumé croît. De manière analogue au cas précédent également, les couvertures obtenues peuvent aussi être très différentes. En revanche, les résumés des épisodes de la série télévisée présentent cette fois-ci des couvertures très élevées, débutant entre 70% et 95%. Le documentaire et l'épisode de fiction présentent quant à eux des couvertures qui varient entre 10% et 25%. Une fois encore, nous expliquons cette attitude par le fait que le documentaire et l'épisode de fiction sont 2 à 3 fois plus longs que les six épisodes

de la série télévisée «Friends»; respectivement le nombre de lignes de la matrice dans ce cas étant égal au nombre d'images représentant chaque vidéo, une très grande diversité survient pour les vidéos plus longues. De plus, la nature de la mise en scène de ces deux vidéos entraîne la présence d'un grand nombre de plans de contenus divers dans ces derniers par rapport aux deux autres vidéos.

4.7 Analyse globale des expériences

Suite aux expériences réalisées, nous constatons que l'application de notre processus de construction automatique de résumés visuels basé sur le PRM produit des résumés ayant de bonnes performances par rapport à notre critère d'évaluation.

Nous ne pouvons faire une comparaison directe des performances obtenues des résumés construits en utilisant les deux représentations étudiées (histogrammes par région et histogrammes de blobs) car les vecteurs caractéristiques des images construits selon ces deux représentations ne sont pas équivalents les uns aux autres. Cependant l'analyse des résultats de l'étude de similarité que nous avons effectuée indique que la représentation des images par des histogrammes de blobs est plus proche du jugement humain de la ressemblance visuelle que la représentation par des histogrammes par région.

Pour la détermination de la similarité visuelle entre deux images, nous avons étudié le cas où une classification stricte basée sur un algorithme de type k-means était utilisée et le cas où une matrice de similarité était réalisée. Intuitivement, l'utilisation de la matrice de similarité est plus proche du raisonnement humain que la l'utilisation de catégories. Dans le cas où une matrice de similarité est employée, une image i peut ressembler à un très grand nombre d'images sans aucune restriction. Par ailleurs, dans le cas où la classification est rigide, une image n'est similaire qu'aux images de sa classe. Aussi, il faut noter qu'au sein d'une même classe, deux images situées à une distance supérieure à la moitié du seuil par rapport au centroïde peuvent avoir une distance entre elles supérieure au seuil de similarité. C'est-à-dire que les images appartenant à la même classe ne sont pas nécessairement évaluées comme similaires en utilisant la matrice de similarité. De même, il est possible de trouver des couples d'images appartenant à deux classes différentes dont la distance est inférieure au seuil de similarité.

En contrepartie, la construction de résumés vidéos sans classification stricte est plus coûteuse en temps de calcul et en espace mémoire, parce que la matrice de similarité qui représente l'ensemble des classes doit être chargée en mémoire durant le processus de sélection d'images représentatives. Par conséquent, l'utilisation de la matrice de similarité n'est pas adaptée à la construction de résumés de vidéos de grande taille.

4.8 Conclusion

Dans ce chapitre, nous avons détaillé notre méthodologie de construction de résumés vidéos, basée sur le PRM en utilisant uniquement les informations visuelles. Nous avons étudié deux différentes représentations des images sous forme de vecteurs caractéristiques. La première utilise des histogrammes de couleurs par région, et la deuxième les histogrammes de blobs. De même, pour la caractérisation de la similarité visuelle, nous avons traité le cas où une classification stricte est réalisée et le cas ou une matrice de similarité est créée. Enfin nous avons comparé et analysé les différents résultats obtenus.

Chapitre 5

Construction de Résumés Multi-vidéos

Dans ce chapitre nous nous intéressons au problème de la construction automatique de résumés de plusieurs vidéos simultanément; une application envisagée est par exemple la construction des résumés de plusieurs épisodes d'une même série télévisée où les mêmes acteurs sont présents dans la plupart des épisodes. La construction de résumés indépendants ne semble pas la meilleure solution.

Dans ce cas de résumés multi-vidéos, nous proposons une extension du principe de reconnaissance maximale permettant de définir les résumés optimaux. Cette extension introduit une nouvelle expérience. Nous ferons ensuite une comparaison de plusieurs variantes de notre méthode, ainsi que des méthodes inspirées par les travaux de deux autres groupes de recherche.

Enfin, nous présentons une expérience basée sur une évaluation réaliste. Des utilisateurs réels réaliserons la tâche de reconnaissance maximale définie dans le chapitre 3. Ils sont pour cela conviés à identifier l'origine d'une centaine d'extraits qui leur sont présentés toute en ne disposant que de l'ensemble des résumés créés par notre approche automatique de création de résumés multi-vidéos basée sur le principe de reconnaissance maximale.

5.1 Résumés Multi-Vidéos

La plupart des travaux actuels se focalisent généralement sur la construction du résumé d'une seule vidéo, seuls quelques uns se sont portés au problème

de résumés multi-vidéos où la prise en compte d'autres contraintes et éléments s'impose. Nous citons par exemple le fait que plusieurs informations (scènes, acteurs) sont présentes telles qu'elles sont ou d'une façon similaire dans diverses vidéos. La construction de résumés vidéos indépendamment les uns des autres peut provoquer une présence d'informations redondantes dans les résumés créés. Des méthodes spécifiques doivent être conçues afin de prendre en considération ces similarités et produire un ensemble de résumés plus efficaces. Ainsi, le résumé de chaque épisode doit contenir les éléments qui caractérisent le mieux cet épisode par rapport aux autres.

Afin de proposer une solution à ce problème, nous présentons une nouvelle approche. Cette dernière consiste à construire simultanément les résumés des différentes vidéos considérées. Ceci en prenant en compte leurs similitudes, de façon à inclure dans chaque résumé les éléments qui différencient le plus la vidéo qui lui correspond par rapport aux autres vidéos traitées.

Ce chapitre est construit de la manière suivante. Nous commençons par la présentation de l'extension de notre principe de reconnaissance maximale au cas des multi-vidéos. Ensuite nous proposons plusieurs sous-variantes de notre algorithme de construction et nous les comparons entre elles ainsi qu'avec deux autres algorithmes inspirés de travaux de recherche existants. Après nous analysons et discutons les résultats des différentes expériences réalisées. Une fois que les résumés sont construits, nous étudions leur robustesse. Enfin nous détaillons l'expérience réaliste d'évaluation de la qualité des résumés construits menée par des utilisateurs réels et nous concluons.

5.2 Principe de Reconnaissance Maximale

Nous proposons une extension du principe de reconnaissance maximale pour la construction de résumés multi-vidéos. Pour ceci, nous adaptons l'expérience que nous avons proposée pour la construction d'un résumé correspondant à une seule vidéo afin de prendre en considération les contraintes engendrées par le cas des résumés multi-vidéos. L'idée intuitive sur laquelle se base notre approche est la suivante: nous désirons construire pour chaque épisode un résumé qui représente le contenu spécifique de cet épisode. Les résumés doivent être créés d'une manière

efficace afin de faciliter la tâche de reconnaissance maximale. Ils sont censés aider les utilisateurs à identifier et distinguer l'épisode duquel provient un extrait pris aléatoirement parmi toutes les vidéos considérées avec le minimum de confusion possible. Ceci nous conduit à proposer le scénario suivant pour cette nouvelle expérience basée sur la reconnaissance maximale illustrée par la figure 5.1:

Figure 5.1: Scénario de l'expérience de reconnaissance maximale appliquée à plusieurs vidéos.

Les trois étapes de l'expérience sont:

- Nous montrons l'ensemble de tous les résumés à l'utilisateur,
- Nous lui présentons un extrait pris aléatoirement d'un flux vidéo quelconque correspondant à une des vidéos traitées,
- Nous lui demandons ensuite, de deviner de quelle vidéo cet extrait provient.

Le comportement simulé de l'utilisateur est le suivant :

1. Si l'extrait contient au moins une image similaire à une ou plusieurs images appartenant à un seul résumé, l'utilisateur indiquera la vidéo correspondante en tant que réponse (notons que cette réponse n'est pas forcément correcte).

2. Si l'extrait comporte des images similaires à d'autres images appartenant à plusieurs résumés, l'utilisateur est dans une situation ambiguë et ne donne pas de réponse.
3. Si l'extrait ne contient aucune image similaire à aucune image de l'ensemble des résumés, l'utilisateur n'a aucune indication et ne donne aucune réponse.

La performance de l'utilisateur lors de cette expérience est le pourcentage de réponses correctes qu'il est capable de fournir en observant tous les extraits possibles tirés de l'ensemble des flux vidéos traités. Notant que parmi les trois cas prédéfinis, c'est seulement au premier que l'utilisateur identifie une vidéo particulière. Cependant, cette réponse n'est pas forcément la bonne, parce qu'une image d'un extrait tiré d'une vidéo peut être similaire à une image incluse dans le résumé d'une autre vidéo.

Cette approche, comme celle proposée dans le cas d'une vidéo unique introduit deux hypothèses: la première est que l'utilisateur dispose d'une mémoire visuelle parfaite, ce qui lui permet d'identifier immédiatement les images similaires lorsqu'elles lui sont montrées. La deuxième hypothèse est qu'on peut automatiquement reproduire d'une manière vraisemblable le jugement de similarité d'un utilisateur réel par le calcul de distance présenté dans le chapitre précédent.

5.2.1 Construction Automatique

Maintenant que nous avons présenté notre objectif de construction simultanée des résumés de plusieurs épisodes d'une même série télévisée, nous formalisons ce problème et nous présentons en premier lieu l'algorithme général utilisé pour la création des résumés multi-vidéos les plus appropriés à la réalisation de notre tâche de reconnaissance maximale. Ensuite nous présenterons des variantes de cette méthode en rajoutant ou éliminant certaines contraintes.

On dénote par E^v un extrait quelconque de la vidéo v, et S_v un résumé de la vidéo v, les cas décrits précédemment peuvent être caractérisés d'une manière formelle par les propriétés suivantes :

- **Le cas Non Ambigu**

  L'extrait E_i^v est identifié comme provenant de la vidéo v'

$$\exists v' \;\; \exists j \quad f_j \in E_i^v \; et \; C(f_j) \in S_{v'}, \;\; \forall v" \neq v' \; \forall f_j \in E_i^v \quad C(f_j) \notin S_{v"} \tag{5.1}$$

- **Le cas Ambigu**

  L'extrait E_i^v peut provenir d'au moins deux vidéos distinctes v' et $v"$

$$\exists v' \exists v" \neq v' \;\; \exists j \;\; \exists k \;\; f_j \in E_i^v \; et \; f_k \in E_i^v \; et \;\; C(f_j) \in S_{v'} \; et \; C(f_k) \in S_{v"} \tag{5.2}$$

- **Le cas inconnu**

  L'extrait E_i^v ne semble parvenir d'aucune vidéo

$$\forall v' \;\; \forall f_j \in E_i^v \quad C(f_j) \notin S_{v'} \tag{5.3}$$

La performance de l'utilisateur est le nombre de réponses correctes, donc c'est le nombre de cas non ambigus pour lesquels $v' = v$:

$$card \left\{ \begin{array}{c} (i,v) : \exists j \;\; f_j \in E_i^v \; et \; C(f_j) \in S_v, \\ \forall v' \neq v \; \forall f_j \in E_i^v \; C(f_j) \notin S_{v'} \end{array} \right\} \tag{5.4}$$

Nous gardons la même définition de similarité d'images que pour les résumés de vidéos uniques, avec la seule différence que pour la construction de résumés multi-vidéos les classes de similarité sont définies d'une manière globale pour l'ensemble des vidéos prises en compte. La construction des résumés devient plus délicate, parce que lorsqu'une classe est sélectionnée pour être rajoutée à un résumé, non seulement la couverture de cette classe dans cette vidéo courante doit être grande mais aussi sa couverture à travers les autres vidéos doit être réduite afin de minimiser le risque d'ambiguïté. La couverture d'une classe dans une vidéo v est définie comme suit :

$$Cov_v(C) = Card\{i : \; \exists j \quad f_j \in E_i^v \;\; et \;\; C(f_j) = C\} \tag{5.5}$$

Une énumération exhaustive de l'ensemble des résumés possibles est presque irréalisable par rapport au temps de calcul. En pratique, nous utilisons un algorithme sous-optimal afin de construire de bons résumés.

Notre algorithme procède comme suit :

1. Chaque résumé est initialement vide.
2. Chaque vidéo v est sélectionnée à tour de rôle, on rajoute à son résumé S_v la classe C qui a la valeur maximale:

$$value_v(C|\{S_v\}) = Cov_v(C|S) - \alpha \sum_{v' \neq v} Cov_{v'}(C|S) \tag{5.6}$$

où S est le groupe des classes déjà incluses dans l'un des résumés : $S = \bigcup_{v'} S_{v'}$

3. Lorsque les différents résumés atteignent la taille désirée, nous remplaçons d'une manière itérative chacune des classes sélectionnées si nous trouvons une autre classe ayant une meilleure valeur.

Le coefficient α est utilisé pour imposer une pénalité aux classes dont la couverture sur les autres vidéos est importante, parce que ces dernières ont tendance à générer des cas d'erreur ou d'ambiguïté.

Il est important de noter qu'à cause du deuxième facteur de l'équation 5.6 $-\alpha \sum_{v' \neq v} Cov_{v'}(C|S)$, nous ne pouvons plus utiliser la majoration uni-vidéo qui permet de simplifier la recherche de la solution optimale en coupant des branches.

5.3 Etapes de création des résumés multi-vidéos

Le procédé de construction de résumés multi-vidéos comprend cinq étapes : pré-traitement du flux vidéo, construction des vecteurs caractéristiques, classification, sélection des segments vidéos et présentation des résumés. Les trois premières ainsi que la dernière sont communes aux six algorithmes que nous présenterons dans la section suivante. La quatrième, qui effectue la sélection des éléments à inclure dans le résumé est quant à elle spécifique à chaque méthode.

5.3.1 Pré-traitement du flux vidéo

Nous traitons la construction de résumés de plusieurs épisodes de la même série télévisée. Malgré que le générique de début et celui de la fin comportent des éléments importants de la série, ils ne sont pas intéressants pour un utilisateur qui essaye de comprendre le contenu d'une vidéo particulière car ils ne spécifient pas un épisode donné. C'est pour ces raisons nous avons décidé d'éliminer les génériques de début et de fin des vidéos utilisées lors de nos expériences.

5.3.2 Construction de vecteurs caractéristiques

Cette étape consiste à analyser le contenu des vidéos pour représenter les données visuelles sous forme de vecteurs caractéristiques. Nous utilisons des histogrammes de blobs 11x11 pour capturer la distribution de couleur ainsi que des informations locales de chaque image. Afin de minimiser le coût de calcul et l'espace mémoire, nous faisons un sous-échantillonnage de la vidéo à une image par seconde. De plus, lorsque des images consécutives sont très similaires, c'est-à-dire lorsque la différence de leurs histogrammes de blobs est inférieure à un seuil prédéfini, nous gardons seulement la première image et nous préservons l'information de durée en incrémentant un compteur associé à l'histogramme.

5.3.3 Classification

Les images représentées par leurs vecteurs caractéristiques sont classifiées par une méthode de k-Means. Dans l'étape d'initialisation on crée une nouvelle classe chaque fois que la distance de l'image par rapport aux classes existantes est supérieure au seuil de similarité que nous avons désigné dans l'étude de similarité visuelle que nous avons effectuée dans le chapitre précédent. Ensuite plusieurs itérations du type k-Means sont réalisées afin d'améliorer le contenu des classes. Cette classification d'images basée sur la comparaison des histogrammes respectifs est supposée produire des classes d'images visuellement similaires.

5.3.4 Sélection des Segments vidéos

Pour chaque épisode, nous sélectionnons les classes caractéristiques les plus importantes en se basant sur les six méthodes différentes qui seront présentées en détail dans la section suivante.

5.3.5 Présentation du résumé

Le résumé global peut être construit et présenté à l'utilisateur sous forme d'une collection d'images représentatives du contenu des vidéos sélectionnées par l'étape précédente ou sous forme d'une séquence audio-visuelle d'une durée réduite obtenue

par la concaténation des segments vidéos correspondant aux classes sélectionnées. Dans ce chapitre, les résumés des divers épisodes sont présentés sous forme d'une grille d'images représentatives extraites des vidéos représentant les classes sélectionnées où chaque ligne représente un épisode donné comme le montre la figure ci-dessous 5.2. Le nombre de lignes dans la grille coïncide avec le nombre d'épisodes pris en considération lors des expériences. Le nombre d'images sélectionnées par épisode (le nombre de colonnes dans la grille) est défini par l'utilisateur.

Figure 5.2: Résumé présenté sous forme d'une collection d'images.

5.4 Les différentes méthodes de sélection

Dans ce qui suit, nous présentons six méthodes ou variantes différentes utilisées pour la construction de résumés multi-vidéos. Un processus automatique représentant un utilisateur simulé est ensuite employé pour évaluer la qualité des différents résumés résultants de ces six méthodes. Ce processus est basé sur notre principe de reconnaissance maximale. Enfin, nous comparons et discutons les résultats d'évaluation afin de définir l'algorithme le plus approprié par rapport à la tâche de reconnaissance maximale.

En premier lieu nous décrivons le principe de quatre algorithmes basés sur notre principe de reconnaissance maximale, ensuite nous expliciterons le reste des algorithmes.

5.4.1 Méthode 1

Dans cette première méthode, un algorithme de base inspiré directement de notre principe de reconnaissance maximale est utilisé pour construire les résumés d'un ensemble d'épisodes de la même série. Cet algorithme s'énonce comme suit: les résumés des épisodes sont construits simultanément. A tour de rôle une classe est sélectionnée et insérée à un résumé correspondant à la vidéo en cours de traitement. La classe sélectionnée est celle ayant la plus grande couverture conditionnelle dans la vidéo considérée sachant l'ensemble des classes insérées dans les différents résumés jusque là. La couverture conditionnelle d'une classe C est égale au nombre d'extraits d'une durée prédéfinie provenant de la vidéo courante contenant la classe C et ne comportant aucune autre classe déjà insérée dans les résumés en cours de construction. Ceci corresond au cas où la valeur zéro est affectée au coefficient $\alpha(\alpha = 0)$ dans l'équation 5.6.

$$value_v(C|\{S_v\}) = Cov_v(C|S) \tag{5.7}$$

où S est le groupe des classes déjà incluses dans l'un des résumés : $S = \bigcup_{v'} S_{v'}$

Lors du calcul de la couverture conditionnelle les extraits contenant des classes déjà sélectionnées sont négligés. Ce qui garantit la sélection d'une seule classe donnée. Elle ne peut pas appartenir à deux résumés différents.

La figure 5.3 représente un exemple illustratif de l'application de cette première méthode de construction de résumés multi-vidéos. Les trois barres à gauche représentent une projection de trois séquences visuelles. Chaque texture forme une substitution d'une succession d'images appartenant à la même classe de similarité. Imaginons que nous désirons construire des résumés dont chacun comporte deux images représentatives afin de synthétiser le contenu de ces trois séquences visuelles. Aussi ces résumés doivent nous permettre de mieux les distinguer entre eux en identifiant facilement et d'une manière si possible certaine l'origine d'un extrait tiré aléatoirement.

Au début les trois résumés sont vides. Notre algorithme commence par sélectionner la première classe à être insérée dans le premier résumé correspondant à la vidéo V_1. Pour ceci, il calcule les couvertures conditionnelles de toutes les classes composant cette première vidéo. A cette étape de la construction les résumés ne

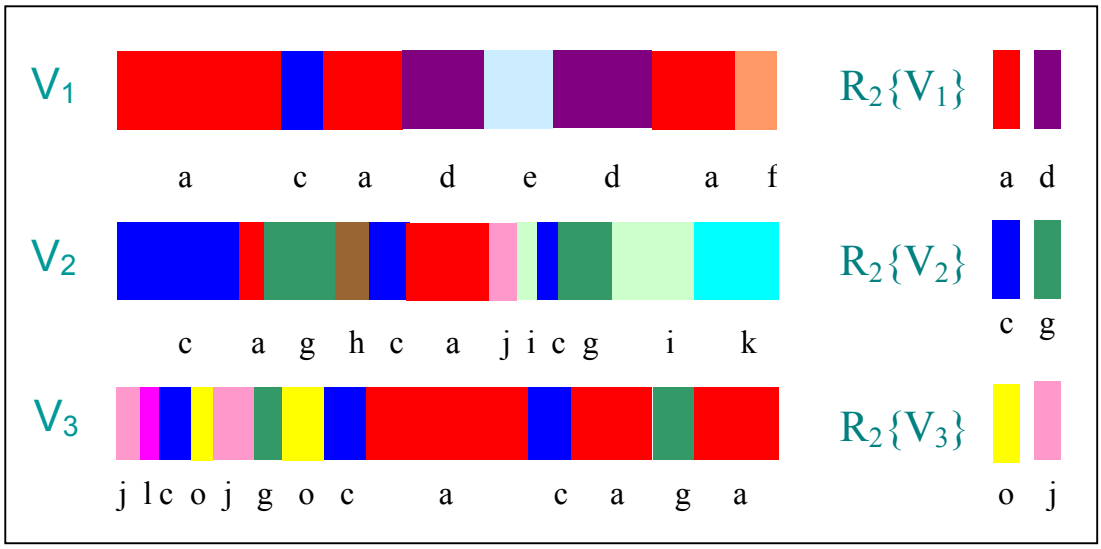

Figure 5.3: Exemple illustratif de la première méthode.

contiennent à présent aucune classe donc les couvertures conditionnelles calculées sont égales aux couvertures des classes, soit, pour une classe donnée, au nombre d'extraits comportant une image de cette classe. Ensuite, il trie les couvertures et sélectionne la classe de premier rang. Cette classe est celle qui représente le mieux le contenu visuel de la première vidéo selon notre principe de reconnaissance maximale, c'est à dire celle qui permet d'identifier le maximum d'extraits tirés de cette première vidéo. L'image représentative de cette classe symbolisée dans notre figure par la lettre (a), sera insérée au résumé de la première vidéo initialement vide.

La prochaine étape de notre processus de construction de résumé multi-vidéos concerne la sélection d'une nouvelle classe afin de l'insérer dans le résumé de la deuxième vidéo. De même que pour la première vidéo, nous calculons les couvertures conditionnelles des classes constituant cette deuxième vidéo sachant qu'une première classe était rajoutée au premier résumé. La couverture conditionnelle de chaque classe présente dans la deuxième séquence vidéo est égale au nombre d'extraits tirés de cette vidéo comportant cette classe et ne comportant pas la classe déjà insérée dans le premier résumé. La classe ayant la plus grande couverture est sélectionnée. Dans cet exemple, c'est une image de la classe (c) qui est insérée dans le deuxième résumé afin de synthétiser le contenu de la deuxième

vidéo et permettre son identification au travers de ses composants.

Enfin, pour la troisième vidéo, les deux classes ayant les plus grandes couvertures sont déjà insérées dans les deux résumés précédents donc leurs couvertures conditionnelles sont nulles et ne peuvent pas être sélectionnées. La classe ayant la plus grande couverture conditionnelle sera celle présente dans un grand nombre d'extraits ne contenant aucune des deux classes sélectionnées auparavant. Dans notre exemple, il s'agit de la classe (o). L'image de cette classe la plus proche du centroïde est insérée dans le troisième résumé jusque là vide. L'algorithme reprend avec la première vidéo et réitère le processus jusqu'à la fin de la construction des trois résumés de taille prédéfinie (2 pour cet exemple).

5.4.2 Méthode 2

Dans la première méthode, la sélection des classes à insérer dans les résumés en cours de construction se base uniquement sur le calcul des couvertures conditionnelles des classes au sein de la vidéo concernée en prenant en compte les classes déjà sélectionnées. Chaque classe sélectionnée ne fait partie que d'un seul résumé à la fois. Imaginons le cas où une classe donnée C est présente dans tous les épisodes traités. Si nous construisons des résumés avec la première méthode, cette classe appartiendra forcement à un résumé parmi ceux créés. Tous les extraits comportant cette classe provenant d'autres vidéos que celle correspondante au résumé contenant cette classe porteront à confusion lors de la phase d'évaluation de la qualité des résumés créés.

Dans le but de diminuer le nombre des cas ambigus ou erronés, nous proposons une variante de la première méthode. Cette deuxième méthode se différencie par la prise en compte, lors de la sélection, des couvertures conditionnelles des classes candidates à travers les autres vidéos que celle en cours est définie comme la «meilleure» classe à insérer à l'étape j au résumé de la vidéo i celle ayant une grande couverture conditionnelle dans la vidéo i ainsi qu'une petite couverture conditionnelle dans les autres vidéos. Cette nouvelle contrainte est formalisée par l'utilisation d'un coefficient négatif en attribuant la valeur de 1 au coefficient $\alpha(\alpha = 1)$ dans l'équation 5.6. Par conséquent, une pénalité est imposée sur les classes ayant une large couverture dans les autres vidéos.

$$value_v(C|\{S_v\}) = Cov_v(C|S) - \sum_{v' \neq v} Cov_{v'}(C|S) \tag{5.8}$$

où S est le groupe des classes déjà incluses dans l'un des résumés : $S = \bigcup_{v'} S_{v'}$

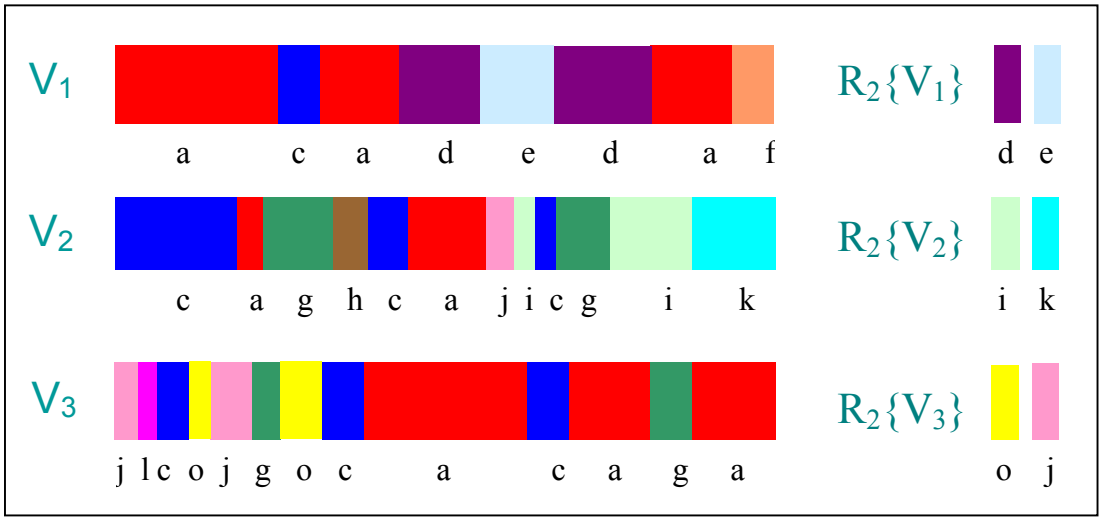

Figure 5.4: Exemple illustratif de la deuxième méthode.

Afin de mieux comprendre le déroulement de ce nouvel algorithme basé sur le principe de reconnaissance maximale avec une contrainte de couvertures inter-vidéos, nous reprenons le même exemple illustratif représenté par la figure 5.4.

Tout d'abord, notre algorithme sélectionne une première classe afin d'être insérée dans le résumé de la première séquence visuelle. La classe choisie est celle qui possède non seulement une grande couverture dans la première vidéo mais aussi une faible couverture dans les deux autres vidéos. La classe symbolisée dans notre figure par la lettre (a) et qui était choisie par la première méthode n'est plus optimale par rapport à ce nouveau critère de sélection. Ceci est dû au fait que cette classe est très présente dans la deuxième et la troisième séquence visuelle. Dans cet exemple, nous remarquons que la classe (d) enregistre une grande couverture conditionnelle dans la vidéo V_1 et n'est pas présente dans les deux autres vidéos, ce qui fait d'elle la meilleure candidate par rapport à notre critère de sélection. C'est cette classe que notre deuxième algorithme sélectionne après avoir calculé les valeurs attribuées aux classes composant cette première vidéo et effectué un tri décroissant. Rappelons que la valeur attribuée à chaque

classe est égale à la couverture conditionnelle de cette classe dans la première vidéo diminuée de la somme de ses couvertures conditionnelles dans les deux autres séquences visuelles. Une fois que cette classe est sélectionnée, une image la représentant est rajoutée au résumé. Cette image est l'image de la vidéo V_1 la plus proche du centroïde de la classe sélectionnée.

Ensuite, le système entame la désignation de la classe la plus «représentative» de la deuxième vidéo par rapport à notre principe de reconnaissance maximale. Cette classe est celle ayant la plus grande couverture conditionnelle dans la deuxième vidéo ainsi que des couvertures conditionnelles minimales dans la première et troisième vidéos. Il s'agit dans cet exemple de la classe (i). La procedure continue ainsi de proche e proche jusqu'à ce que le système ait construit la totalité des résumés.

5.4.3 Méthode 3

Dans les deux premières méthodes, une classe sélectionnée une première fois pour faire partie d'un résumé correspondant à une vidéo donnée ne peut être sélectionnée une deuxième fois. Elle ne peut apparaître deux fois dans le résumé de la même classe ni dans deux résumés différents. Ce principe est assuré par notre définition de la couverture conditionnelle de chaque classe. Nous considérons la couverture conditionnelle d'une classe C lors du calcul du résumé de la vidéo i, le nombre d'extraits tirés de la même vidéo i comportant cette classe mais surtout ne comportant aucune classe insérée jusque là dans l'ensemble des résumés sans aucune exception. Le fait d'interdire à une classe d'appartenir à plusieurs résumés crée une dépendance et un lien étroit entre les divers résumés. Cette façon de faire nous permet de diminuer le taux de confusion et engendre des résumés plus spécifiques et plus distinctifs de l'ensemble des vidéos traitées.

Afin de valider les atouts de la prise en considération de l'ensemble des vidéos lors du calcul d'un résumé particulier dans le cadre global de construction multi-vidéos, nous proposons une troisième variante où la construction de chaque résumé se fait indépendamment des autres.

Pour pouvoir comparer les différentes variantes de notre méthode de base inspirée du principe de reconnaissance maximale, nous utilisons toujours une

classification globale. Ce qui implique qu'avant la phase de sélection dont les algorithmes diffèrent, nous effectuons une classification de toutes les images composant les différents épisodes traités sous forme de classes de similarité.

Dans cette troisième méthode la définition de la couverture conditionnelle diffère de celle des deux premières méthodes. Elle s'énonce comme suit: «La couverture conditionnelle d'une classe C appartenant à une vidéo est égale au nombre d'extraits tirés de cette vidéo comportant cette classe et ne contenant aucune autre classe déjà insérée au résumé correspondant à cette même vidéo». Il découle de cette définition, que les extraits contenant la classe C et une autre classe déjà sélectionnée mais insérée dans un autre résumé que de celui en cours sont préservés. Ainsi une classe appartenant à un autre résumé peut donc être sélectionnée pour la deuxième fois ou plus mais cette fois-ci pour faire partie du résumé courant.

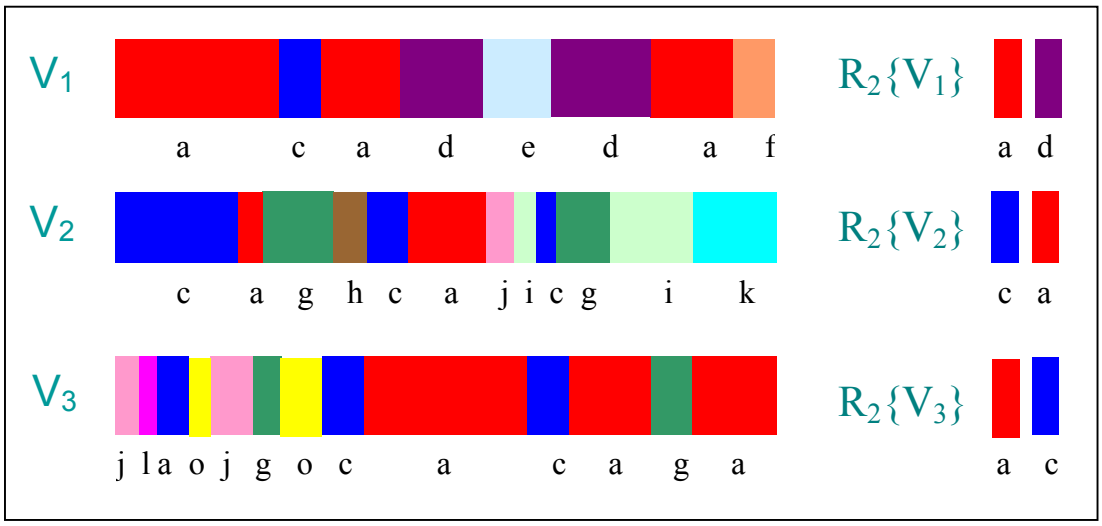

Figure 5.5: Exemple illustratif de la troisième méthode.

Afin d'expliquer d'avantage l'idée, nous utilisons toujours l'exemple de la figure 5.5. En récapitulant, cette troisième méthode de sélection permet la construction du résumé de chaque vidéo indépendamment des autres vidéos. Il n'est plus indispensable de procéder à tour de rôle car l'ordre dans lequel sont traitées les vidéos devient insignifiant. Chaque vidéo est considérée comme unique et c'est exactement le processus de construction de simple résumé présenté dans le

chapitre précédent qui est employé.

Nous observons par exemple pour la première vidéo V_1 qu'il n'y a pas de lien de voisinage étroit entre les deux classes ayant les plus grandes couvertures symbolisées par la lettre (a) et la lettre (d). Ces deux classes seront sélectionnées après le calcul des couvertures conditionnelles et insérées dans le résumé correspondant à cette vidéo. Il en est de même pour les deux autres vidéos.

Par ailleurs, nous remarquons que la classe (a) est présente d'une manière considérable avec une couverture assez importante dans les trois vidéos. Ainsi cette classe sélectionnée va être incluse dans chacun des trois résumés créés. Pour le premier résumé, nous substituons cette classe par l'image la plus proche du centroïde originaire de la première vidéo. Quant au deuxième résumé, nous insérons une image de la même classe mais originaire de la deuxième vidéo. Enfin pour le dernier résumé, nous rajoutons une troisième image de la même classe mais cette fois ci, prise dans la troisième vidéo. Donc les différents résumés construits, ne peuvant pas contenir des images identiques, peuvent contenir des images similaires.

5.4.4 Méthode 4

Nous avons proposé dans la deuxième méthode d'imposer une pénalité sur les classes ayant une large récurrence dans les différentes vidéos afin de diminuer les cas ambigus et erronés. Nous proposons maintenant, une nouvelle variante de notre méthode de base qui permet d'éliminer d'une manière absolue tous les cas ambigus.

Dans notre expérience de reconnaissance maximale, nous retrouvons une ambiguïté lorsqu'un extrait quelconque contient soit une image similaire à deux images appartenant à deux résumés distincts, soit au moins deux images dont chacune est similaire à une image présente dans un résumé différent de l'autre. Ces deux cas peuvent être présentés autrement: le premier consiste à avoir une même classe insérée dans plusieurs résumés à la fois, pour le deuxième, les classes ne sont pas redondantes mais il existe au moins deux classes incluses dans deux résumés distincts dont les images comprises en elles sont voisines dans la limite de la taille de l'extrait utilisé dans au moins une séquence originale.

Afin d'éliminer tous les cas ambigus dans notre expérience simulée, nous développons un algorithme basé sur le calcul de la couverture de la même façon que dans les méthodes précédentes, sauf que les classes candidates ne doivent pas être présentes ni dans les autres résumés ni dans les extraits qui contiennent des classes déjà sélectionnées dans les résumés des autres vidéos.

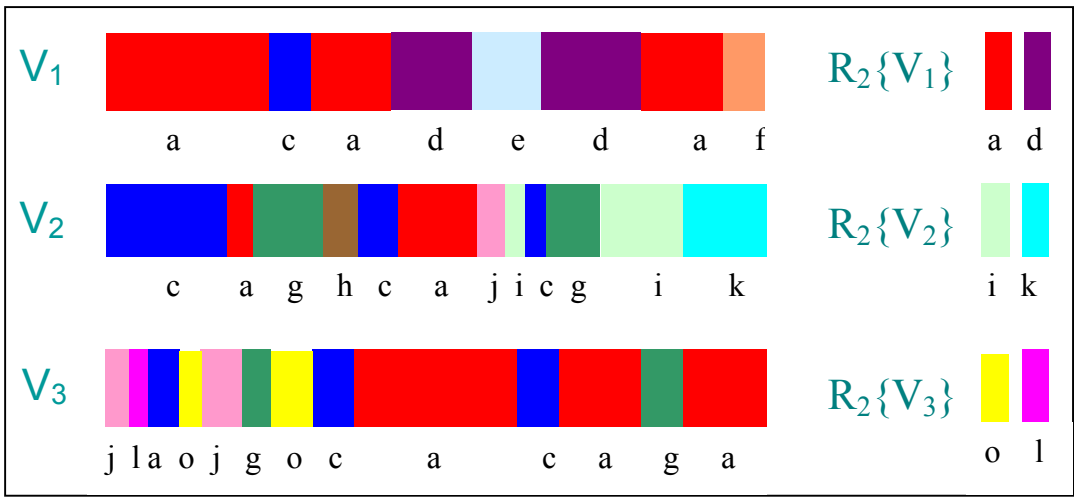

Figure 5.6: Exemple illustratif de la quatrième méthode.

Nous gardons toujours, le même exemple illustratif de la figure 5.6 pour expliquer cette quatrième méthode de sélection. L'algorithme débute par la sélection de la première classe représentative de la première vidéo. Initialement les résumés sont vides, donc il ajoute au premier résumé la classe ayant la plus grande couverture dans la première vidéo quelle que soit sa couverture dans les autres vidéos. A ce niveau de la construction il n'y a pas de risque d'ambiguïté parce que c'est la première classe sélectionnée. Dans ce cas de figure, c'est la classe (a) qui est la plus présente dans la première séquence visuelle. Donc l'image la plus proche de son centroïde provenant de cette première vidéo est insérée dans le premier résumé.

Ensuite, le système procède à la sélection d'une nouvelle classe à inclure dans le résumé de la deuxième vidéo. Pour ceci il commence par le calcul des couvertures conditionnelles des classes composant cette dernière. De façon analogue à la première et la deuxième méthode, nous définissons la couverture conditionnelle d'une classe appartenant à une vidéo donnée comme le nombre d'extraits

comportant cette classe et ne contenant aucune autre classe déjà comprise dans l'ensemble des résumés en cours de construction. Cette définition nous garantit la suppression de la première cause d'ambiguïté c'est-à-dire, l'apparition de la même classe dans deux résumés différents. Dans notre exemple, il est admis que lors du calcul des couvertures conditionnelles des classes appartenant à la deuxième vidéo, tous les extraits comportant une image de la classe (a) insérée dans le premier résumé sont ignorés. Par conséquent, les classes qui prennent des valeurs de couvertures conditionnelles non nulles ne sont certainement pas voisines des classes du premier résumé (elles n'appartiennent à aucun extrait comportant une classe du premier résumé). Ces classes sont triées en fonction de leur couverture conditionnelle et traitées dans l'ordre décroissant. La première qui remplit la nouvelle condition que nous lui imposons s'assurer de l'élimination de la deuxième cause d'ambiguïté est sélectionnée. Elle sera rajoutée dans le résumé de la deuxième vidéo V_2. La condition consiste à ne jamais être présente dans un extrait quelconque des autres vidéos (V_1, V_3) comportant la classe insérée dans le premier résumé. Nous remarquons que la classe (c) est celle qui possède la plus grande couverture conditionnelle mais ne peut être sélectionnée parce qu'elle est voisine de la classe (a) dans les trois vidéos. Il y aura donc au moins un extrait de chaque vidéo comportant ces deux classes. Ceci qui implique directement la génération d'un cas d'ambiguïté où il est impossible de deviner de quelle vidéo provient chacun de ces extraits, (de la première, de la deuxième vidéo ou tous simplement de la troisième?) c'est la confusion totale. Il en est de même, pour la classe ayant la deuxième plus grande couverture (g). Par ailleurs, la classe possédant la troisième couverture en ordre décroissant (i) n'est pas voisine de la classe existante dans le premier résumé. C'est l'image provenant de la deuxième vidéo la plus proche du centroïde de cette classe qui sera rajoutée au deuxième résumé.

Pour la troisième vidéo, il faut choisir la classe ayant la plus grande couverture dans cette vidéo et qui ne soit ni sélectionnée auparavant ni présente conjointement dans le même extrait que les deux classes déjà insérées dans le premier et le deuxième résumé en cours. Dans cet exemple c'est la classe (o) qui remplit ces conditions et qui donc est sélectionnée et substituée par une image la représentant dans le troisième résumé. Une fois que le système a fait un premier tour de

l'ensemble des vidéos, il reprend le même processus afin d'augmenter la taille de chaque résumé.

Il faut juste noter qu'une classe candidate d'une vidéo donnée peut être voisine d'une ou de plusieurs classes appartenant au résumé de la même vidéo. Dans cet exemple, la deuxième classe à rajouter au résumé de la première vidéo indiquée par la lettre (d) est voisine de la première classe du résumé correspondant à celle indiquée par la la lettre (a). Cependant, elle n'est surtout pas voisine d'aucune des deux autres classes insérées à ce niveau aux résumés des deux autres vidéos. L'addition de cette nouvelle classe ne crée aucun cas d'ambiguïté, au contraire les extraits contenant des images des deux classes du premier résumé seront plus indicatifs et plus faciles à deviner que les autres extraits.

5.4.5 Méthode 5

Uchihachi et Foote [UF99], ont défini une mesure pour le calcul de l'importance des plans, et nous adaptons cette mesure à notre méthode de construction de résumés multi-vidéos. Dans la méthode originale, la classification hiérarchique mise en œuvre commence par l'attribution de chaque image à une classe unique. Ensuite, le nombre de classes est réduit en regroupant à chaque étape d'une manière itérative les deux classes les plus proches. Ce regroupement est basé sur le calcul de la distance minimale entre toutes les combinaisons possibles de deux classes d'images. Chaque feuille de cet arbre dont la racine constitue une seule classe comportant toutes les images (la vidéo en entier), représente un singleton comprenant une seule image. Chaque niveau de l'arborescence correspond à une classification en un nombre variable de classes. Une fois que la classification des images composant la vidéo (ou une partie des images dans le cas du sous-échantillonnage) est réalisée, ces auteurs ont défini une mesure de calcul d'un facteur d'importance attribué à chaque plan. Pour ceci, ils ont affecté à chaque classe un poids normalisé W_i qui est défini comme étant la proportion de plans parmi tous ceux de la vidéo et qui appartiennent à la classe i. La formule de calcul de ce poids est la suivante:

$$W_i = \frac{S_i}{\sum_{j=1}^{C} S_j} \tag{5.9}$$

où C est le nombre de classes considéré et correspondant à un niveau de la hiérarchie. Ces classes comportent toutes les images de vidéo prise en compte et S_i est la longueur totale de l'ensemble des plans appartenant à la classe i, obtenue par la sommation des longueurs de tous les plans appartenant à cette classe.

Ils considèrent un plan comme étant important si ce dernier représente les deux caractéristiques suivantes: la longue durée et la rareté (ne ressemble pas à la plupart des autres plans). Suite à cette définition le facteur d'importance d'un plan est mesuré en combinant la longueur du plan avec le poids inverse de la classe à laquelle il appartient. Donc l'importance I du plan j (de la classe k) est calculée comme suit:

$$I_j = L_j \ Log\frac{1}{W_k} \tag{5.10}$$

où L_j est la longueur du plan j.

A la différence de la méthode originale, nous effectuons une classification de type k-means au lieu d'une classification hiérarchique. Nos segments (appelés plans) sont construits après la phase de classification de l'ensemble des images des épisodes traités gardées après la phase de sous-échantillonnage et d'élimination de génériques. Les plans sont construits en concaténant les images successives appartenant à la même classe. Aussi, au lieux d'une seule vidéo, nous considérons un certain nombre des épisodes.

Dans notre étude, afin de représenter chaque vidéo par des plans spécifiques et les plus longs possibles, nous calculons le facteur d'importance pour tous les plans possibles. Ensuite nous sélectionnons dans chaque vidéo les plans les plus importants afin de les inclure dans les résumés correspondants.

5.4.6 Méthode 6

L'idée principale de cette méthode est de faire un parallèle avec les méthodologies de construction de résumés texte, où la formule $TF.IDF$ a prouvé qu'elle est

très efficace pour définir l'importance d'un mot du texte. Cette formule combine la fréquence du terme tf avec la fréquence inverse du document idf. Le tf mesure la densité d'un terme (mot) dans un document donné. Le idf mesure l'utilité d'un mot, c'est-à-dire si c'est un mot commun ou non à l'ensemble des documents du corpus. Le idf peut être calculé uniquement en utilisant le nombre de documents comportant une occurrence de ce mot ($idf = 1/df$). La version utilisée généralement est celle où on considère le LOG de la fréquence inverse du document idf. Pour les résumés de texte, les unités de base sont les mots alors que pour les résumés multi-vidéos les unités sont les classes d'images similaires. Par analogie avec la formule $TF.IDF$, nous définissons l'importance I de la classe c de la manière suivante:

$$I_c = L_c \log n/n(c) \tag{5.11}$$

où L_c est la longueur (durée totale) de la classe c, n le nombre de vidéos et $n(c)$ le nombre de vidéos contenant au moins une image de la classe c. Ayant calculé l'importance de chaque classe, nous sélectionnons les plus importantes qui composeront le résumé global. Dans le cas où une classe est présente dans plusieurs vidéos, nous devons déterminer à quel résumé nous l'affectons. Nous faisons ceci en calculant pour chaque vidéo la proportion d'images appartenant à cette classe présentes dans la vidéo, et en choisissant la plus probable.

5.5 Expériences

Dans cette section nous présentons les résultats d'évaluation en utilisant le principe de reconnaissance maximale sur les résumés vidéos construits avec six différentes sélections par le biais de notre utilisateur simulé qui reproduit le comportement d'un utilisateur réel. Nous avons effectué une série d'expériences avec des vidéos de type $Mpeg1$. Ces vidéos représentent six épisodes de la série télévisée «Friends» qui ont été enregistrés à 14 $images/sec$ (c'est une limite de la carte d'acquisition). Nous avons choisi de présenter nos résumés sous la forme d'une grille avec six images par épisode. Ceci est particulièrement adapté à l'affichage sur un écran de télévision ou d'un ordinateur.

La figure 5.7 illustre le résultat de notre première méthode de création de résumés multi-vidéos. Chaque ligne de cette grille d'image est spécifique à une des six vidéos.

Figure 5.7: Résumé des six épisodes "Friends" construits par notre première méthode de sélection.

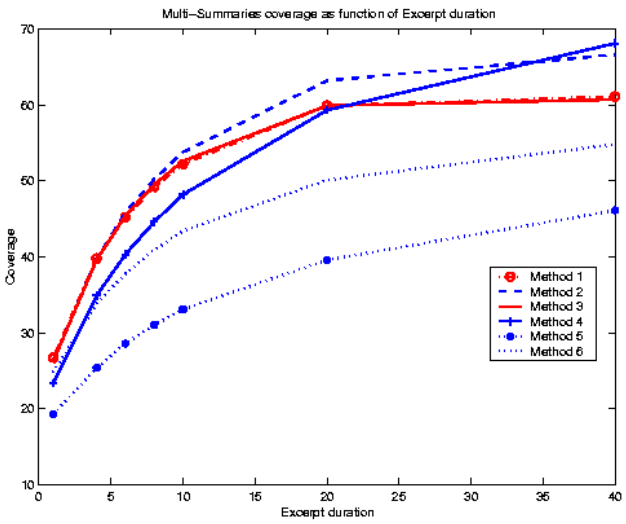

Figure 5.8: Performances des résumés des six épisodes en fonction de la méthode de sélection.

Le graphe de la figure 5.8 présente les performances respectives des six méthodes de sélection utilisées quand la durée de l'extrait utilisé lors de l'évaluation varie. Nous notons que les deux premières méthodes qui construisent les résumés en se basant sur un critère mathématique inspiré du critère d'évaluation lui même donnent les meilleurs performances. Nous notons aussi que les résumés multi-épisodes (méthodes 1 et 2) sont plus efficaces que les résumés de vidéos simples où chacun est construit indépendamment des autres (méthode 3). La cinquième méthode consiste à sélectionner les plans les plus longs et les plus rares. Malgré le fait que notre critère de performance est basé sur le taux de réponses correctes (les extraits pour lesquels lors de l'évaluation les vidéos originales respectives sont bien devinées) et le fait que les plans rares sont par définition très spécifiques par rapport aux vidéos desquelles ils sont tirés, les résultats ne sont pas bons comparés aux autres méthodes. Ceci est dû au fait que les plans rares ont tendance à ne pas être redondants dans la vidéo (leur nombre d'occurrence est très faible souvent égal à un), donc certainement quelle que soit leur longueur, ils ne s'étendent pas tout au long de la vidéo et ont une petite couverture à travers cette dernière. Ces caractéristiques provoquent une réduction significative du nombre d'extraits

possibles comportant des images originaires des plans sélectionnés, et par conséquent sur les performances qui sont obtenues. La méthode 6, inspirée de la formule $TF.IDF$ donne des résultats moyens par rapport aux autres. Nous notons aussi que les résultats de la quatrième méthode sont comparables à ceux de la deuxième, et que les deux donnent les meilleures couvertures pour des extraits de longue durée.

Nous observons dans la figure 5.8 que pour une durée d'extrait approximative 30 secondes les résultats de la deuxième méthode (avec $\alpha = 1$) et la quatrième méthode (sans ambiguïté) sont meilleurs que les résultats de la méthode basique (méthode 1 avec $\alpha = 0$), cependant c'est le contraire pour les extraits plus courts. Une explication possible consiste à considérer le fait que pour les extraits longs, la probabilité de trouver des images similaires dans d'autres vidéos augmente. Dans ce cas de figure la réponse a plus tendance à être fausse ou indéterminée (cas ambigu), ce qui implique forcément une réduction de la performance. Cela n'arrive pas avec les deux autres méthodes quelle que soit la durée de l'extrait, car les deux principes de sélection prennent respectivement en considération la diminution et l'élimination des cas ambigus. De même, la troisième méthode procure des performances raisonnables pour les extraits courts. Chaque fois que la durée de l'extrait utilisé pour la construction de résumés augmente, le nombre de cas ambigus augmente aussi. Ceci est dû au fait que des images de la même classe de similarité peuvent appartenir à plusieurs résumés parce qu'une même classe peut être attribuée à plus d'un résumé à la fois.

5.6 Robustesse des résumés

Ayant construit des résumés multi-vidéos en utilisant un certain nombre de méthodes, il est intéressant d'évaluer la performance des résumés pour une durée d'extrait donnée. Les quatre premières méthodes sont dépendantes de la durée d'extrait cependant les deux dernières ne le sont pas. Afin d'étudier la robustesse des résumés, nous construisons des résumés multi-vidéos avec une certaine durée d'extraits et ensuite nous les évaluons avec d'autres durées. La Figure 5.9 montre les résultats de cette expérience pour des résumés construits avec la première méthode (avec $\alpha = 0$). Nous avons choisi de faire cette étude en utilisant la

		Méth-ode1	Méth-ode2	Méth-ode3	Méth-ode4	Méth-ode5	Méth-ode6
Durée des extraits	1	26.7%	26.3%	26.3%	23.4%	19.3%	24.8%
	4	39.8%	39.7%	39.6%	34.9%	25.4%	33.8%
	6	45.2%	45.8%	45.4%	40.3%	28.6%	37.8%
	8	49.2%	50.2%	49.5%	44.6%	31.0%	40.9%
	10	52.2%	53.8%	52.6%	48.2%	33.1%	43.4%
	20	59.9%	63.2%	59.9%	59.3%	39.6%	50.1%
	40	61.1%	66.6%	60.7%	68.1%	46.1%	54.8%

Tableau 5.1: Résultats de performances des résumés pour les six méthodes sous considération(blob seuil=455)

première méthode parce que lors des expériences de calcul de performance, cette dernière a donné en général les meilleurs performances. Les résultats présentés dans le tableau 5.2 montrent que pour chacun des résumés la meilleure couverture est obtenue lorsque nous utilisons, lors de l'évaluation, des extraits de longue durée. D'après ce graphe, nous observons qu'à l'exception des résumés construits avec une durée d'extrait égale à 1 seconde, les autres donnent des performances similaires et de bon niveau. La couverture croît de la même façon pour tous les résumés, ce qui indique que pour des durées d'extraits moyennes (ni trop grandes ni trop petites) utilisées lors de la création, les méthodes donnent des résumés robustes par rapport aux conditions d'évaluation.

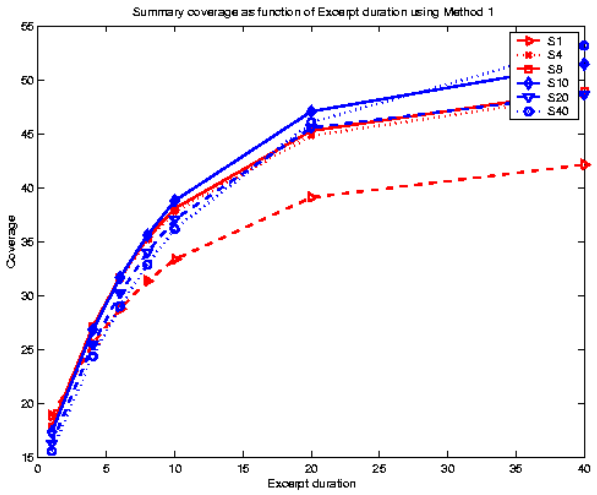

Figure 5.9: Etude de la robustesse des résumés.

		Durée d'extrait lors de la construction						
		1	4	6	8	10	20	40
Durée d'extrait lors de l'évaluation	1	18.9%	18.0%	17.8%	17.8%	17.4%	16.3%	15.6%
	4	25.6%	27.1%	27.1%	27.1%	26.8%	25.5%	24.4%
	6	28.8%	31.6%	31.7%	31.7%	31.7%	30.2%	29.0%
	8	31.4%	35.2%	35.3%	35.3%	35.6%	34.0%	32.9%
	10	33.4%	37.8%	38.1%	38.1%	38.8%	37.1%	36.2%
	20	39.1%	44.8%	45.3%	45.3%	47.1%	45.6%	46.1%
	40	42.1%	48.6%	49.0%	49.0%	51.5%	48.7%	53.2%

Tableau 5.2: Résultats de performance des résumés utilisant la première méthode avec plusieurs durées d'extraits utilisés lors de la construction et l'évaluation

5.6.1 Temps de Calcul

Bien que nous n'ayons pas optimisé notre code, nous fournissons quelques éléments d'information sur la complexité de nos algorithmes. Les temps d'exécution reportés dans cette partie ont été obtenus lors de l'exécution du processus complet de la construction automatique des résumés multi-vidéos comme décrit précédemment sur une station de travail SUN Ultra10. Il faut 5 minutes pour calculer et

sauvegarder les 2705 histogrammes représentant l'ensemble des 99 minutes de vidéos (les six épisodes de «Friends»). La classification des images représentatives en 1285 classes dure environ 3 heures et 30 minutes.

La méthode de sélection utilisée est celle de base (méthode 1 avec $\alpha = 0$), et le temps de calcul requis pour différentes durées d'extraits est présenté dans le tableau 5.3 ci-dessous. Il est intéressant de noter qu'avec l'implémentation du mécanisme de sélection fondé sur un principe de type «greedy» le temps de calcul croît d'une façon sous-linéaire. Finalement, l'évaluation est extrêmement rapide et ne nécessite que quelques secondes pour l'estimation des performances d'un résumé.

Malgré le fait que nous ne construisons pas de résumés vidéos automatiques en temps réel, nous pensons qu'avec l'utilisation d'une optimisation appropriée ainsi que des outils hardware la construction des résumés multi-vidéos peut être effectuée dans une durée équivalente à celle de la vidéo.

	Durée	des	extraits	en phase	de	const	ruction
	1	4	6	8	10	20	40
Temps de calcul	47sec	50sec	53sec	56sec	58sec	71sec	97sec

Tableau 5.3: Temps de calcul du processus de sélection

5.7 Evaluation des utilisateurs réels

Jusque là nous avons en premier construit différents résumés selon diverses méthodes de sélection. Ensuite nous les avons évalués à l'aide de notre utilisateur simulé fondé sur des hypothèses sur le comportement des utilisateurs réels. Dans le but de valider ces hypothèses ainsi que de mesurer l'écart du comportement simulé par rapport à la réalité, nous avons organisé une expérience basée sur le principe de reconnaissance maximale avec des utilisateurs réels. Un autre objectif de cette expérience consiste à faire une analyse du comportement des utilisateurs réels afin de trouver des moyens nous permettant d'améliorer les performances de notre méthodologie de construction et d'évaluation automatique.

L'ordre de présentation des différentes expériences dans ce mémoire ne correspond pas exactement à l'ordre chronologique dans lequel nous les avons réalisées.

Au début, nous avons construit les résumés singuliers et les résumés multi-vidéos en utilisant les histogrammes de régions de couleurs comme étant des vecteurs caractéristiques des images traitées. Ensuite nous avons introduit une nouvelle représentation qui consiste aux histogrammes de blobs. Cette expérience avec les utilisateurs réels était réalisée avant l'introduction de la nouvelle caractérisation c'est à dire les blobs. Donc les résumés utilisés ont été construits avec les histogrammes de régions et les performances du système (ou plutôt de l'utilisateur simulé) ont été aussi calculées à base des histogrammes de régions.

Le scénario de l'expérience est le suivant :

- Un utilisateur réel regarde l'ensemble des résumés comme présenté dans la figure 5.10 où chaque ligne représente le résumé d'un épisode spécifique.

Figure 5.10: Résumé présenté aux utilisateurs réels.

- Nous montrons ensuite à l'utilisateur un extrait tiré des vidéos, et il essaye de deviner de quelle vidéo cet extrait était pris. L'utilisateur peut revoir l'extrait autant de fois qu'il veut, pour pouvoir comparer les images de l'extrait avec le résumé présenté, cela nous permet de vérifier l'hypothèse que l'utilisateur possède une mémoire visuelle parfaite, cf. figure 5.11.

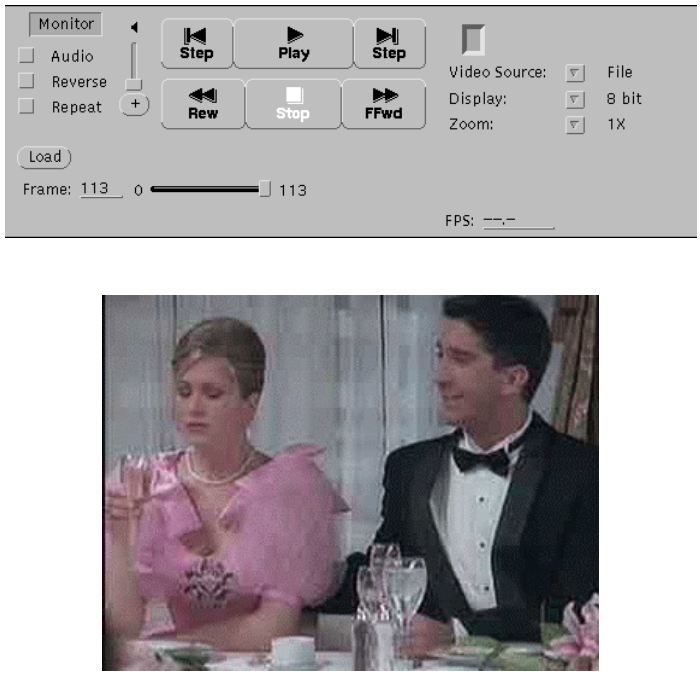

Figure 5.11: Présentation à l'utilisateur d'un extrait tiré aléatoirement.

Dans notre expérience nous avons montré à chaque utilisateur 100 extraits tirés aléatoirement des six vidéos étudiés. Au fur et à mesure de l'avancement de l'expérience, nous comptons le nombre le réponses correctes et de réponses fausses et nous discutons avec chaque utilisateur des raisons de son choix. Le procédé est répété pour dix utilisateurs distincts (membres de notre institut) qui participent à cette expérience. Comme présenté dans le tableau 5.4, tandis que la performance de notre système automatique est de 36%, les utilisateurs réels enregistrent une performance moyenne de 82.9% (avec des valeurs minimale et maximale de 79% et 89%). Les utilisateurs réels ont des meilleurs performances parce qu'ils sont capables d'analyser le contenu des images et de concevoir une représentation sémantique des scènes et des évènements existant dans la vidéo (il est intéressant de noter que les utilisateurs qui réalisent les meilleurs performances sont des fans des séries télévisés en particulier «Friends»). Une explication des performances établies par notre système est la simplicité de notre mesure de similarité d'images. Comme inconvénient de cette mesure nous pouvons citer le fait que cette représentation n'est pas invariante aux transformations d'images

suivantes: Translation/Rotation (les mêmes acteurs peuvent être vus à travers différents angles de caméra) et l'Echelle (un acteur peut occuper presque toute l'image ou seulement une partie). Finalement, il faut attirer l'attention sur le fait que malgré que notre système a des performances réduites lors de l'évaluation du résumé construit, ceci n'implique pas que notre résumé est mauvais, par contre les performances des utilisateurs réels montrent le contraire, c'est à dire avec ce résumé un utilisateur quelconque peut reconnaître en moyenne 82.9% des extraits qui lui sont montrés.

Utilisateurs	Système	Moyenne des performances des utilisateurs réels	MIN	MAX
Performance	36%	82.9%	79%	89%

Tableau 5.4: Comparaison des performances des utilisateurs réels et l'utilisateur simulé

5.8 Analyses des résultats

Le tableau 5.5 reporte le nombre d'extraits devinés correctement par n utilisateurs (maximum $n = 10$) et le nombre de ceux-ci qui ont été devinés correctement par notre système de construction basé sur le principe de reconnaissance maximale. Afin d'analyser ces résultats davantage, nous considérons les deux premiers cas, pour lesquels tous ou neuf sur dix des utilisateurs réels ont donné les réponses correctes. Il y a 73 extraits $(61 + 12)$ pour lesquels le système ne procure que 26 bonnes affectations $(19+7)$. Comme notre objectif est d'améliorer la performance de notre méthodologie de construction automatique de résumés vidéos, nous avons étudié plus précisément les 47 cas parmi les 73 pour lesquels le système a échoué.

Pour chacun de ces 47 cas, nous cherchons les raisons probables qui ont permis aux utilisateurs réels d'identifier d'une manière appropriée les vidéos originales de ces séquences visuelles. Suite aux commentaires faits par les différents utilisateurs lors de la réalisation de cette expérience, cinq facteurs majeurs sont pris en considération: Personne, Objet, Action, Emplacement, temps. Ces cinq facteurs sont ensuite affinés sous forme de neuf critères; un seul acteur et ses habits, différents acteurs et leurs habits, des habits spécifiques (cravate, etc...), des détails

Nombre d'utilisateurs répondant correctement	Nombre des extraits	Nombre des extraits bien devinés par le système
10	61	19
9	12	7
8	6	4
7	4	2
6	3	1
5	1	1
4	2	0
3	0	0
2	0	0
1	4	1
0	6	0

Tableau 5.5: Comportement des utilisateurs réels et de l'utilisateur simulé

de visage inhabituels, un acteur occasionnel, des objets inhabituels, une action en cours, le décors, des détails remarqués dans une image et retrouvés dans d'autres images.

Ayant défini ces neuf critères, nous avons regardé chaque extrait et affecté un pourcentage d'importance à chaque critère indiquant le degré d'utilité de chacun d'eux dans la prise de décision concernant la bonne vidéo. Cette affectation était faite manuellement d'une façon entièrement subjective. Malgré que cette affectation est subjective, elle préserve de l'information utile.

Le tableau 5.6 résume en premier temps nos résultats du calcul des pourcentages d'importance moyens pour chaque critère en prenant en compte l'ensemble des 47 extraits. A partir de ces données, nous constatons que le critère «les acteurs et leurs habits» est le critère le plus fréquemment utilisé pour reconnaître la vidéo correspondante à l'extrait, suivi de près du critère «un acteur et ses habits». Ensuite, le nombre de fois où un seul critère est décisif (degré d'importance supérieur à 50 pour cent) est reporté. Ceci nous permet, en combinant avec le pourcentage moyen d'importance, de retrouver l'importance d'un

critère particulier.

	Moyenne(47 extraits)	Nb extraits ≥50%
Acteur + Habits	25.96	7
Acteurs + Habits	27.45	9
Habits Spécifiques	16.59	7
Détails de Visage	3.4	3
Acteur.Occasionnel	3.4	0
ObjetNon habituel	2.34	0
Action	10.32	1
Décors	4.47	0
1/n images	7.55	0

Tableau 5.6: Pourcentage d'importance de chaque critère de décision

Il est important de projeter nos résultats sur les perspectives du domaine de traitement d'images actuel et les recherches de la vision par ordinateur. En effet, il est intéressant de savoir quelles sont les méthodologies qui permettraient d'améliorer la qualité des résumés et de rapprocher les performances de l'utilisateur simulé de celle de l'utilisateur réel. En se basant sur les neuf critères que nous avons définis, nous avons déduit cinq fonctions de traitement d'images potentiellement utiles.

- **Fonction 1:** Reconnaissance et détection de visages.
- **Fonction 2:** Reconnaissance et détection des objets.
- **Fonction 3:** Identification d'une personne ainsi que ses habits (reconnaissance et détection de visage + reconnaissance et détection d'habits).
- **Fonction 4:** Identification d'un groupe de personnes avec leurs habits.
- **Fonction 5:** Reconnaissance d'objets et de personnes dans des environnements distincts.

Nous pouvons calculer le nombre des extraits pour lesquels l'association avec la bonne vidéo peut être effectuée. Pour chacune des fonctions, nous faisons l'hypothèse qu'il existe un algorithme qui la réalise d'une manière robuste et parfaite. Le tableau 5.7 représente nos résultats et montre que les 47 extraits

peuvent être bien classifiés en devinant leurs vidéos respectives si une méthode d'identification d'une personne complète avec ses habits est disponible.

	Fonct1	Fonct2	Fonct3	Fonct4	Fonct5
Nb des extraits que le système peut deviner correctement	23	18	47	32	34

Tableau 5.7: Le nombre d'extraits qui peuvent être rattrapés avec l'utilisation de chacune des fonctions proposées

En conséquence, l'analyse basée sur la reconnaissance et la détection du visage et d'un modèle du corps semble la fonction la plus prometteuse pour l'amélioration des performances. Notons toutefois que les valeurs précédentes sont certainement très spécifiques au type des vidéos utilisées lors des expériences.

5.9 Conclusion

Une comparaison de plusieurs variantes de construction automatique de résumés multi-vidéos a été présentée. En se basant sur le principe de reconnaissance maximale, nous évaluons les résultats obtenus par six méthodes différentes. Nos expériences démontrent que les meilleurs résultats sont réalisés lorsque la construction et l'évaluation sont effectuées avec le même principe. La méthode que nous avons proposée donne clairement de meilleurs résultats que les méthodes inspirées d'autres travaux existants. Notre évaluation de la robustesse des résumés montre qu'il est possible d'obtenir de bonnes performances avec des résumés construits en utilisant des durées moyennes d'extraits.

Notre évaluation avec l'intervention d'utilisateurs réels nous a permis de valider la qualité de notre résumé et nous a aussi aidé à faire une comparaison entre l'évaluation d'utilisateurs simulés et réels. Enfin cette étude nous a orienté vers les fonctions de traitement d'images qui permettront dans le futur d'améliorer les performances de notre système.

Chapitre 6

Construction de Résumé par le Texte

Ce chapitre présente une extension de notre approche prenant en compte la transcription textuelle du flux audio. Cette transcription peut également être remplacée par les sous-titres diffusés au travers du télétexte. Nous étudions le cas où la transcription est prise en considération toute seule comme étant un document texte isolé ainsi que le cas où elle fera partie d'un contexte global; ce qui correspond à la présence d'un corpus incluant le document texte traité. Pour les deux éventualités, nous construisons des résumés textuels composés d'un certain nombre de mots sélectionnés suivant notre PRM.

6.1 Introduction

Comme nous l'avons présenté dans l'état de l'art, plusieurs travaux ont été réalisés dans le domaine de la construction de résumés textuels. D'après Mani et al. [MM99] ces approches peuvent être classifiées en différentes catégories: les méthodes basées sur des approches classiques comme les approches statistiques [Luh58] [Edm69], les méthodes basées sur l'utilisation de corpus de documents [KPC95] [cAGLO99], ainsi que des approches qui exploitent la structure du discours de la langue utilisée pour la rédaction de ces documents [BK97] [BE97] ou qui utilisent des connaissances de haut niveau comme les informations sémantiques [Leh81] [May95]. Le nombre de méthodes proposées dans la littérature est important et le choix d'une de ces méthodes dépend des besoins et de l'utilité du résumé

construit. Dans notre cas, nous désirons construire un résumé textuel de façon à pouvoir le combiner avec le résumé visuel que nous avons construit, afin de finalement obtenir un résumé Multimédia (Visuel et Textuel). Nous souhaitons rester cohérent par rapport à notre méthode de construction de résumés visuels. Dans ce but nous proposons une méthode de construction de résumés textuels basée sur le principe de reconnaissance Maximale exposé dans le chapitre 3 [MHYS02].

6.2 Reconnaissance textuelle maximale

Nous décrivons une nouvelle proposition, qui consiste à utiliser le même principe de reconnaissance maximale PRM pour les informations textuelles. Nous faisons l'hypothèse qu'une transcription de la vidéo est disponible et synchronisée avec la vidéo (elle peut être obtenue à travers une transcription manuelle, extraction de télétexte ou reconnaissance de la parole). Le résumé textuel R^T est composé d'un ensemble de mots ou de phrases. Un extrait E^T est défini comme étant une sous-phrase ou simplement une succession de mots tirée de la transcription originale. Afin d'appliquer le PRM, nous devons définir une règle de décision qui peut être utilisée pour valider si un extrait provient du document qui correspond au résumé ou non. Le scénario de l'expérience que nous proposons est le suivant:

- Un résumé textuel R^T est présenté à l'utilisateur sous forme d'un ensemble de mots.
- Un extrait E^T composé d'une suite de mots tirés de la transcription textuelle lui est ensuite montré.
- L'utilisateur est sollicité de deviner si l'extrait qui lui était présenté provient de la transcription correspondante au résumé R^T.

Nous proposons trois variantes de la règle de décision suivant lesquelles le comportement de l'utilisateur diffère. Ces trois variantes représentent des degrés différents concernant l'évaluation de l'évidence du bon choix du résumé. L'évaluation peut être souple, intermédiaire ou stricte.

Dans ces règles de décision, l'utilisateur décide qu'un extrait provient du document original:

- Si au moins un mot de l'extrait E^T appartient au résumé R^T (politique souple).
- Si tous les mots de l'extrait E^T appartiennent au résumé R^T (politique stricte).
- Si au moins un pourcentage de mots P de l'extrait E^T sont présent dans le résumé R^T (politique intermédiaire).

La première règle de décision correspondante à la politique souple est très similaire à la règle de décision utilisée pour le processus de construction des résumés mono-vidéo. Alors que cette décision apparaît parfaitement valide pour la vidéo parce que les images identiques sont rares, elle est moins correcte pour le texte parce que l'occurrence d'un mot donné n'est pas nécessairement une indication forte de similarité. C'est la raison pour laquelle nous avons défini les deux autres politiques: intermédiaire et stricte.

Quelle que soit la règle de décision utilisée, la probabilité des réponses correctes désigne la performance enregistrée de l'utilisateur lors de cette expérience. En fonction de la politique (la règle de décision utilisée) nous définissons plus explicitement la performance de la sorte:

- Dans le cas de la politique souple, la performance est égale au nombre d'extraits textuels E^T tirés de la transcription, associé à la vidéo originale, comportant au moins un mot inséré dans le résumé textuel R^T.
- Tandis que dans le cas de la politique intermédiaire, la performance est le nombre moyen des extraits qui contiennent un pourcentage de mots P participant au résumé.
- Enfin pour la politique stricte, la performance est le nombre moyen des extraits qui ne contiennent que des mots appartenant au résumé.

6.3 Construction du résumé textuel

Ayant exprimé clairement le principe de reconnaissance textuelle maximale, nous présentons maintenant la méthodologie de construction de résumé textuel d'une manière formelle. Rappelons que le facteur principal de la sélection des mots composant le résumé est la règle de décision. De ce fait, pour chaque politique

mentionnée précédemment, une règle de décision doit être définie. Cependant, la méthodologie de base reste la même. Notre objectif est de maximiser la performance des mots sélectionnés pour constituer le résumé R de la même façon que le cas où seules des informations visuelles étaient prises en considération.

Pour la construction du résumé texte, nous commençons notre processus par une phase préliminaire de pré-traitement dans laquelle les mots communs appartenant à des «*stop-lists*» prédéfinies sont exclus du processus de sélection d'une façon analogue aux approches utilisées dans le domaine de la recherche d'information. Tous les mots communs de ces listes présents dans la transcription sont tout simplement ignorés parce que ces derniers n'apportent pas d'informations significatives et spécifiques au document traité.

Ensuite, nous appliquons une analyse morphologique (stemming) dans laquelle les racines $R(M)$ des différents mots M sont calculées. Les mots considérés sont ceux de la transcription associée au document multimédia considéré, gardés après l'élimination des mots outils lors de la phase préliminaire.

Après le calcul des racines, nous classifions ces derniers sous forme de classes de similarité W où chaque classe comporte des mots ayant la même racine (même morphème). Dans notre approche, nous considérons deux mots M_i et M_j similaires si et seulement si ils ont des racines identiques:

$$M_i \textit{ est similaire à } M_j \iff W(M_i) = W(M_j) \iff R(M_i) = R(M_j) \quad (6.1)$$

Le résumé texte optimal de taille k peut être retrouvé en énumérant tous les ensembles possibles contenant k classes de mots $\{W_1, W_2, ...W_k\}$ et en conservant celui qui maximise la performance moyenne de tous les extraits possibles E du document D. L'énumération de la totalité des ensembles semble très coûteuse en temps de calcul. De même que dans le cas des résumés des informations visuelles, il est judicieux et plus rentable de sélectionner minutieusement l'ordre dans lequel les classes sont sélectionnées, ainsi la meilleure solution est retrouvée rapidement. En pratique, nous avons utilisé un algorithme basé sur un principe de type «greedy» pour la sélection des unités constituant le résumé. Ce type d'algorithme nous permet de diminuer considérablement le temps d'attente de l'utilisateur et d'avoir des temps de construction raisonnables même si la solution

obtenue est sous-optimale. Avec cette approche, la performance est décomposée par la formule suivante:

$$\begin{aligned} \text{perf}\,(R^T) &= \text{perf}\,(W_1, W_2, ..., W_k) \\ &= \text{perf}\,(W_1, W_2, ..., W_{k-1}) + \text{perf}\,(W_k | W_1, W_2, ..., W_{k-1}) \end{aligned} \tag{6.2}$$

L'algorithme de construction du résumé sous-optimal procède comme suit:

- **Etape 1:** Commencer avec un résumé textuel R^T vide.
- **Etape 2:** Trier les classes qui n'ont pas encore été sélectionnées en fonction de la valeur décroissante de la performance perf(R^T) par rapport à la constitution actuelle du résumé.
- **Etape 3:** Rajouter la classe W ayant une performance maximale pour le résumé. Revenir à l'étape 2 jusqu'à ce que le résumé ait la taille voulue.
- **Etape 4(optionnelle):** Pour raffiner le résumé, prendre chaque unité du résumé à tour de rôle et essayer d'identifier une classe (un autre mot) qui améliore la performance. Cette étape est répétée jusqu'à ce qu'aucune amélioration supplémentaire ne puisse être effectuée. Le résultat final de cette étape consiste en la solution optimale par rapport à notre principe de reconnaissance maximale.

L'algorithme présenté débute par la sélection la classe W_1 ayant une valeur de performance maximale perf(W_1), puis W_2 qui a une valeur de performance maximale perf(W_1, W_2), et ainsi de suite jusqu'à W_k. La première solution complète retrouvée est alors le résultat d'une série de choix avec un critère de type «Greedy». Si nous désirons nous contenter d'une solution sous-optimale en gagnant en temps de calcul, l'étape 4 de notre algorithme de sélection peut être omise.

Cette procédure s'applique telle quelle pour la politique souple. Pour les politiques intermédiaire et stricte, il faut noter que la performance du premier mot à sélectionner est égale à zéro (à moins que les extraits aient une longueur de 1 mot). Donc, nous remplaçons dans ces cas la performance exacte perf(R^T) par une proportion, qui dépend linéairement du nombre des mots de l'extrait existant

déjà dans le résumé. Ceci permet de sélectionner les mots les plus prometteurs, même pour le premier choix de M_1.

Une fois que le meilleur ensemble des classes de mots est trouvé, chaque classe W_i est remplacée par un mot représentatif M_i, ce qui définit l'ensemble des mots qui composent le résumé. Pour le choix du représentant de la classe, nous calculons pour chaque mot son nombre d'occurrences dans le document original, ensuite nous les trions et gardons celui qui se classe au premier rang c'est-à-dire celui qui est le plus présent dans la classe.

Comme indiqué précédemment, le calcul de la fonction de performance perf(R^T) dépend de la politique utilisée pour la création et l'évaluation du résumé. Plus particulièrement, c'est la règle de décision $d(E, R)$ qui est élaborée spécialement pour chacune des trois politiques envisagées.

Lors de la présentation du principe de reconnaissance textuelle, nous avons défini la performance comme étant le nombre de réponses correctes données par l'utilisateur. Nous rappelons aussi que la réponse (décision: l'extrait appartient ou non au document correspondant au résumé) de l'utilisateur se fait selon la règle de décision déterminée. Cette dernière était définie selon les différentes politiques d'évaluation de la pertinence et de la détermination d'un mot dans le contexte de reconnaissance maximale. Dans ce qui suit, nous présentons d'une manière formelle les règles de décision associées aux trois politiques énoncées précédemment.

6.3.1 La politique souple

L'idée motrice de la politique souple consiste à donner aux mots une valeur d'authenticité leur permettant d'être déterministes. Dans ce cas de figure, la présence d'un seul mot du résumé dans un extrait est amplement suffisante pour valider son appartenance au document original du résumé présenté.

La règle de décision qui transpose cette politique souple est formalisée comme suit:

$$d_1^T(E^T, R^T) = \begin{cases} 1 \quad \text{si} \quad \exists M_i \in E^T \quad M_i \in R^T \\ 0 \quad \text{sinon} \end{cases} \tag{6.3}$$

Notre processus de construction est basé sur la notion de classes de mots.

Les mots ayant la même racine sont considérés comme similaires. Nous adaptons notre règle de décision pour que l'utilisateur décide de l'appartenance de l'extrait à la transcription originale du résumé si ce dernier comporte au moins un mot similaire à un ou plusieurs mots du résumé (ayant la même racine).

$$d_1^T(E^T, R^T) = \begin{cases} 1 & \text{si } \exists M_i \in E^T \quad \exists M_j \in R^T \text{ et } W(M_i) = W(M_j) \\ 0 & \text{sinon} \end{cases} \tag{6.4}$$

Cette première règle de décision encourage une grande complémentarité et peu de redondances entre les mots composant le résumé. Ces effets sont dus à notre processus de construction dans lequel nous rajoutons au résumé, à chaque étape de la phase de sélection, le mot représentant de la classe qui permettra d'augmenter au maximum la performance du résumé courant. Dans ce premier cas la performance du résumé courant est égale à la somme des couvertures conditionnelles des classes insérées dans ce dernier. Notons que c'est la classe obtenant la meilleure couverture conditionnelle connaissant les classes insérées dans le résumé parmi les classes non sélectionnées jusque là qui sera insérée dans le résumé en cours de construction et ainsi de suite jusqu'à ce que la taille désirée du résumé soit atteinte. D'une manière analogue aux classes des images, nous définissons la couverture conditionnelle d'une classe de mots comme étant sa contribution à la couverture de l'ensemble des classes composant le résumé courant c'est-à-dire le nombre d'extraits reconnus uniquement grâce à cette classe sans aucune implication des autres classes déjà insérées dans le résumé. En d'autres termes, c'est le nombre d'extraits comportant au moins un mot de cette classe et aucun autre mot appartenant à l'une des classes sélectionnées auparavant.

$$\begin{aligned} Cov(W_m|W_1W_2...W_{m-1}) &= Cov(W_1W_2...W_m) - Cov(W_1W_2...W_{m-1}) \\ &= Card\left\{ \begin{array}{l} i : \exists j \ M_j \in E_i \ \ et \ \ W(M_j) = W_m \\ et \ \ \forall M \in E_i \forall j = 1, 2, ...m-1 \ \ W(M) \neq W_j \end{array} \right\} \end{aligned} \tag{6.5}$$

6.3.2 La politique stricte

Dans le cas de la politique stricte, nous estimons qu'un mot seul n'est pas suffisant pour reconnaître son origine car il peut faire partie de plusieurs documents à la fois. Sa présence conjointe dans l'extrait et le résumé n'est pas une preuve pour valider que l'extrait présenté est tiré du document correspondant au résumé. Cependant la présence de tous les mots composant l'extrait sans exception peut être une attestation de l'origine de cet extrait. Nous considérons cette présence collective comme une affirmation du fait que l'extrait est tiré du document original correspondant au résumé montré.

La règle de décision qui manifeste ce comportement est la suivante:

$$d_2^T(E^T, R^T) = \begin{cases} 1 & si \quad \forall M \in E^T \quad M_i \in R^T \\ 0 & \text{sinon} \end{cases} \tag{6.6}$$

Comme nous utilisons des classes de similarité, notre règle de décision sera comme suit: «L'utilisateur décide de l'appartenance de l'extrait au document original du résumé si et seulement si tous les mots le composant sont similaires à des mots présents dans le résumé.»

$$d_2^T(E^T, R^T) = \begin{cases} 1 & si \quad \forall M \in E^T \quad W(M_i) \in R^T \\ 0 & \text{sinon} \end{cases} \tag{6.7}$$

Lors de la phase de sélection, la performance du résumé courant est égale au nombre moyen des extraits pour lesquels n'importe quel mot M_i est similaire à un mot du résumé.

6.3.3 Politique Intermédiaire

Afin d'avoir un compromis entre les deux politiques souple et stricte, nous proposons une politique intermédiaire, dans laquelle, nous estimons qu'un groupe de n mots peut être déterministe pour la reconnaissance de son origine. La valeur de n dépend de notre tolérance de jugement et de la valeur intrinsèque que nous affectons aux mots. Comme nous utilisons des extraits de diverses tailles où le nombre de mots inclus dans l'extrait diffère d'un extrait à l'autre, nous avons décidé de remplacer le nombre de mots n nécessaires pour la reconnaissance de

l'extrait par un pourcentage P de mots nécessaires par rapport au nombre total de mots composant l'extrait. Avec cette politique, l'utilisateur décide qu'un extrait provient du document original du résumé lui est présenté si un pourcentage de mots P de l'extrait sont présents dans le résumé. Sachant que nous avons élaboré une classification des mots considérés en un ensemble de classes de similarité selon leurs racines, notre règle de décision sera la suivante: «L'utilisateur devine correctement l'origine d'un extrait de mots si ce dernier contient un pourcentage de mots P similaires à des mots du résumé lui est montré». La formalisation de cette règle de décision est la suivante:

$$d_3^T(E^T, R^T) = \begin{cases} 1 & si \quad (NBM/NME) * 100 \geq P \\ 0 & \text{sinon} \end{cases} \tag{6.8}$$

où $NBM = card(M \in E^T \text{ tel que } W(M) \in R^T)$

et NME : Nombre de Mots composant l'Extrait (taille de l'extrait)

Pendant la construction du résumé le plus adapté à cette politique intermédiaire, et à chaque étape d'insertion d'une classe au résumé courant, nous sélectionnons celle qui permet de maximiser la performance du résumé actuel plus cette classe. La performance du sous-ensemble de classes en question est égale au nombre d'extraits du document original qui contiennent un pourcentage P de mots de ces classes sélectionnées.

Comme nous le constatons, le cas de la politique stricte est un cas particulier de la politique intermédiaire, pour lequel le pourcentage de mots qui doivent être similaires aux mots du résumé est égal à 100%.

Dans le cas de la politique souple, la construction du résumé textuel selon le PRM est basée sur le calcul des couvertures conditionnelles. A chaque étape de sélection nous rajoutons au résumé courant la classe ayant la meilleure couverture. Cependant, dans le cas de la politique intermédiaire (de même que stricte), cette mesure n'est pas envisageable car il faut noter que la performance du premier mot à sélectionner est égale à zéro (à moins que les extraits aient une longueur de 1 mot). Donc, nous remplaçons dans ce cas la performance exacte par une proportion, qui dépend linéairement du nombre des mots de l'extrait existant déjà dans le résumé. Ceci permet de sélectionner les mots les plus prometteurs, même pour le premier choix de M_1. Cette nouvelle mesure pour chaque classe

non sélectionnée dans le résumé courant est formalisée comme suit:

$$Cov(W) = \sum_{NE(W)} (Cov1 + 1/NME) \tag{6.9}$$

où $Cov1 = card(M \in E^T$ tel que $W(M) \in R^T)$
et $NE(W)$= Nombre d'extraits qui contiennent un mot de la classe W

6.4 Les expériences

Dans le but d'expérimenter notre principe de reconnaissance maximale appliqué aux informations textuelles pour la construction des résumés textes, nous avons créé manuellement la transcription du flux audio d'un documentaire intitulé «Histoire d'eau» qui fait partie d'un corpus de vidéos distribué par l'INA (Institut National d'Audio-Visuel). Cette vidéo était utilisée dans les expériences du chapitre 4, où le principe était appliqué aux informations visuelles. La transcription associée à cette vidéo est composée de 4852 mots. Après la phase de pré-traitement qui consiste à éliminer les mots outils (mots communs présents dans les «*stop-lists*»), nous obtenons un document texte renfermant 1813 mots dont 1113 mots distincts. Nous avons ensuite appliqué un procédé d'étude morphologique pour le calcul des morphèmes des mots gardés après la suppression des mots communs. Le nombre de morphèmes distincts obtenu (nombre de classes de mots) est égal à 949.

Une fois que les phases de pré-traitement, d'étude morphologique et de classification étaient établies, nous avons entamé la phase de sélection des mots selon les différentes politiques. Nous avons construit des résumés textuels de taille égale à vingt (chaque résumé contient 20 mots) utilisant différentes longueurs d'extrait (de un à huit mots). Dans un premier temps, nous avons créé des résumés sous-optimaux par rapport à la politique souple où la mesure de la performance du résumé est égale à la somme des couvertures conditionnelles des classes le composant. Ensuite, nous avons utilisé l'heuristique proposée pour construire des résumés qui correspondent conjointement aux politiques stricte et intermédiaire.

La figure 6.1 montre un exemple de résumé textuel (composé de vingt mots)

Petit – Consommation – Goût – Appelle – Gratuit – Construit – Couper – Cuisiner – Travail
An – Quotidien – Analyser – Déverser – Bonne – Protéger – Tournée – Changer
Semaine – Réserves – Ressources

Figure 6.1: Exemple d'un résumé textuel construit selon la politique souple avec des extraits comportant quatre mots.

construit avec la première méthode basée sur les couvertures conditionnelles en utilisant des extraits de quatre mots. Ce résumé est le résultat d'un processus de type «greedy», il est sous-optimal. Par ailleurs, la figure 6.2 représente un résumé, ayant la même taille que le précédent, construit avec l'heuristique proposée pour les cas de la politique stricte est intermédiaire. La longueur des extraits utilisés est égale à quatre mots.

Petit – Consommation – Goût – Tournée – Appelle – Réservoir – Parisiens – Changer
Construit – Protéger – Métier – Conduites – Immeubles – Début – Enterrer – Températures
Quotidien – Bonne – Alimenter – Utilise

Figure 6.2: Exemple d'un résumé textuel construit selon la politique intermédiaire avec des extraits comportant quatre mots.

Une fois que les résumés sous-optimaux étaient construits en utilisant deux mesures différentes, nous les avons évalués selon les trois politiques proposées. Les résumés résultant de notre méthode basée sur le PRM et la mesure de couverture conditionnelle étaient évalués uniquement selon la politique souple. Cependant, les résumés résultant de notre PRM combiné avec la nouvelle mesure proposée au lieu de la couverture conditionnelle étaient évalués selon la politique stricte et la politique intermédiaire en testant différentes valeurs pour le pourcentage des mots qui doivent être présents conjointement dans l'extrait et le résumé afin de valider l'origine de l'extrait.

Le tableau suivant 6.1 représente les performances calculées suite à cette évaluation. Bien sûr, pour le cas simple d'extraits de longueur égale à un, les performances des trois politiques sont identiques. Lorsque la longueur de l'extrait augmente, la performance de la politique souple croît rapidement, parce que de

plus en plus d'extraits contiennent au moins un mot du résumé. Au contraire, la performance de la politique stricte ainsi que l'intermédiaire décroît, car il est plus difficile de trouver des extraits dont l'ensemble ou une grande partie des mots appartiennent au résumé. Notons que les fluctuations des résultats de performance de la politique stricte sont dues à l'heuristique utilisée pour le choix de classes en combinaison avec un procédé de type «greedy», permettant un optimum local. Dans l'expérience de la politique intermédiaire, la variation de la performance est aussi due à l'effet de l'arrondi du nombre minimum de mots requis.

Longueur de l'extrait	Politique Souple d_1^T	Politique Intermédiaire d_3^T $n >= 50\%$	Politique Intermédiaire d_3^T $n >= 75\%$	Politique Stricte d_2^T
1	16%	16%	16%	16%
2	28.7%	27.6%	4.1%	4.1%
3	40.1%	9.4%	1.6%	1.6%
4	49.4%	13.4%	3.6%	1.1%
5	57.5%	4.5%	1.8%	1.2%
6	64.3%	4.3%	1.7%	1.4%
8	75.2%	2.8%	1.6%	1.2%

Tableau 6.1: Performance du résumé(construction et évaluation sans utilisation d'information de contexte)

6.5 Construction de résumés contextuels

Jusque là, notre méthode de construction de résumés textuels basée sur le principe de reconnaissance maximale, quelle que soit la mesure utilisée, ne prenait en compte que le document traité. Il est certain que les mots sélectionnés proviennent du document courant mais rien ne prouve que ces mots sont vraiment spécifiques aux sujets abordés par ce document ou qu'ils sont assez discriminants pour le représenter d'une manière qui permet de distinguer son résumé parmi ceux d'autres documents. Les mots sélectionnés peuvent être fréquents dans ce document sans être forcement particuliers à ce dernier, mais plutôt communs

à plusieurs autres documents. Bien que nous ayons utilisé une «*stop-list*» afin d'éliminer les mots outils les plus communs (articles, prépositions, interjections, etc...), nous n'avons pas remédié complètement à cet effet. Un grand nombre de mots (verbes, noms, adjectifs, etc...) peuvent être communs à divers documents textes.

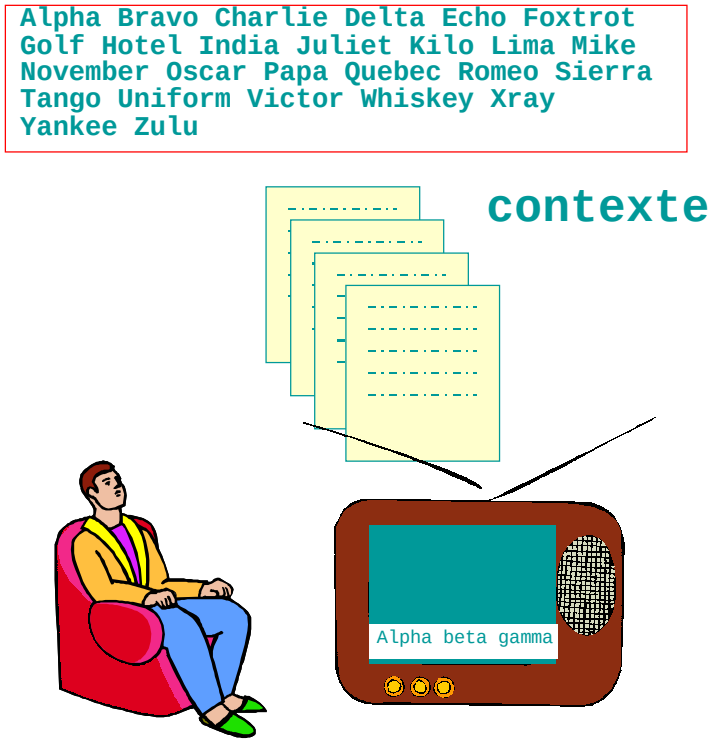

Figure 6.3: Utilisation du contexte pour la construction du résumé.

Les travaux de recherches dans le domaine de la recherche d'information ont introduit des mesures tel que la formule $TF.IDF$ pour combiner les effets de la fréquence du mot et sa puissance discriminative. Afin de transposer ceci vers notre problème de reconnaissance maximale, nous avons considéré que, dans la règle de décision pour identifier un extrait, les mots qui apparaissent fréquemment dans les autres documents doivent être pénalisés. Nous pouvons facilement adapter notre approche pour incorporer cette nouvelle contrainte. Par exemple, nous faisons l'hypothèse qu'une librairie de vidéos personnelles contient différentes vidéos, chacune avec son texte associé (une transcription de l'audio ou un regroupement du télé texte), cf. figure 6.3. Afin d'éviter les ambiguïtés, il faut inclure dans les

résumés de ces vidéos des mots qui ne soient pas présents dans les transcriptions des autres vidéos (ou au moins ils ne soient pas présents dans leurs résumés). Pour la sélection d'un mot pour un résumé spécifique, nous devons prendre en compte la probabilité que ce mot apparaisse dans un autre résumé. Lorsque nous construisons un résumé spécifique, nous dénommons «contexte» l'ensemble des textes existant dans la librairie.

Pour le flux vidéo, nous avons déjà proposé dans le troisième et le quatrième chapitres des méthodologies de construction de résumé simple ainsi que des résumés multi-vidéos. Pour les résumés textuels, il est important pour le résumé d'exclure les mots communs (les mots présents dans le texte courant ainsi que le contexte) et de retenir des mots spécifiques au contenu particulier de ce document par rapport aux autres.

Dans le but d'estimer la probabilité qu'un mot particulier appartienne au contexte, nous considérons comme contexte un ensemble de r documents sélectionnés aléatoirement du Web. Cette probabilité peut alors être estimée à partir des indications fournies par un moteur de recherche sur Internet. Pour mettre en œuvre cette expérience, nous avons utilisé les indications procurées par altavista (www.altavisa.com) concernant la fréquence des mots dans le web et le nombre de documents qui les contiennent. Alors, nous calculons la probabilité que le contexte ne contienne pas le mot M_i suivant les étapes suivantes:

1. Si un mot M_i est présent dans un certain nombre de documents $ND(M_i)$ alors la probabilité $P(M_i \in doc)$ qu'un document aléatoire contienne le mot M_i est égale à:

$$P(M_i \in doc) = ND(M_i)/ND \tag{6.10}$$

   où ND est la taille de notre corpus de documents (la partie du Web indexée par Altavista) et $ND(M_i)$ est le nombre de documents du corpus qui comportent au moins une occurrence du mot M_i (comme indiqué par Altavista dans la page résultante après la recherche du mot M_i).
2. Suite à cette première définition, la probabilité qu'un document aléatoire ne contienne pas le mot M_i est égale à:

$$P(M_i \notin doc) = (ND - ND(M_i))/ND \tag{6.11}$$

3. Enfin, la probabilité que r documents aléatoires (que nous appelons contexte) ne contiennent pas le mot M_i est définie comme suit:

$$P(M_i \notin contexte) = (ND - ND(M_i)/ND)^r \tag{6.12}$$

En introduisant la notion de classes de similarité de mots nous obtenons:

$$P(W_i(M) \notin contexte) = (ND - ND(W_i(M))/ND)^r \tag{6.13}$$

$ND(W_i(M))$ est le nombre de documents du corpus contenant au moins un mot de la classe $W_i(M)$.

La prise en compte d'un contexte comportant des documents supplémentaires par rapport au document traité demande la définition de nouvelles règles de décision à partir des anciennes. Ces nouvelles règles sont probabilistes. Avec la présence du contexte, l'utilisateur prend une décision avec une certaine probabilité d'exactitude. Si un mot est très présent dans les documents constituant le contexte, ce mot ne permettra pas de prendre une décision avec une grande certitude. Donc l'utilisateur est dans l'incapacité d'identifier exactement l'origine de l'extrait mais il décide de son appartenance avec un certain degré de certitude (la probabilité associé à ce choix).

Maintenant, nous réécrivons chacune des règles de décision qui correspondent aux trois politiques pour refléter l'effet de l'utilisation de l'information contextuelle. Nous définissons aussi la performance associée à chaque politique.

6.5.1 La politique souple avec contexte

Dans le cas de la politique souple, la règle de décision est la suivante: «L'utilisateur devine l'origine d'un extrait donné, s'il y a au moins un mot de ce dernier qui est similaire à un mot du résumé et qui n'est pas présent dans le contexte (des r documents aléatoires)». La probabilité que l'utilisateur aie raison est égale à la probabilité que ce mot ou un autre qui lui est similaire n'appartiennent pas à d'autres documents du contexte considéré.

La règle de décision est formalisée comme suit:

$$d_4^T(E^T, R^T) = \begin{cases} \max P(W(M_i)) & si \quad \exists M_i \in E^T \quad W(M_i) \in R^T \\ 0 & \text{sinon} \end{cases} \tag{6.14}$$

Selon cette règle et lors de la construction du résumé les mots ambigus (qui ont une forte probabilité d'apparaître dans le contexte) auront de faibles chances d'être sélectionnés, parce qu'ils contribuent moins à la performance du résumé. La performance du résumé construit sera égale à la somme des probabilités des classes le composant pour tous les extraits auxquels l'une de ces classes appartient.

6.5.2 La politique stricte avec contexte

Pour la politique stricte, la règle de décision est modifiée pour prendre en compte les informations contextuelles de la manière suivante: «L'utilisateur reconnaît l'origine d'un extrait donné, si tous les mots composant ce dernier sont similaires à des mots présents dans le résumé et ne sont pas présents dans le contexte». La probabilité que l'utilisateur ait raison est égale au produit des probabilités que les mots constituant l'extrait ou d'autres qui leur sont similaires n'appartiennent pas à des documents du contexte pris en considération.

$$d_5^T(E^T, R^T) = \begin{cases} P(W) & si \quad \forall M_i \in E^T \quad W(M_i) \in R^T \\ 0 & \text{sinon} \end{cases} \tag{6.15}$$

où $P(W) = \prod_{W_i \in E^T} P(W_i \notin contexte) = \prod_{W_i \in E^T} (ND - ND(W_i)/ND)^r$

La performance du résumé construit selon cette règle est égale à la somme, sur tous les extraits comportant exclusivement des classes du résumé, des produits des probabilités de ces classes à l'intérieur de chaque extrait.

6.5.3 La politique intermédiaire avec contexte

Finalement, la règle de décision qui correspond à la politique souple combinée avec l'utilisation des informations contextuelles est la suivante: «L'utilisateur devine l'origine d'un extrait qui lui est présenté si ce dernier comporte un pourcentage de mots similaires aux mots du résumé sans être présent dans le contexte (des r

documents aléatoires)». La probabilité de cette décision est égale la probabilité que les mots du résumé n'appartiennent pas au contexte.

$$d_6^T(E^T, R^T) = \begin{cases} P(W) & si \quad (NBM/NME) * 100 \geq P \\ 0 & \text{sinon} \end{cases} \tag{6.16}$$

où $NBM = card(M \in E^T \text{ tel que } W(M) \in R^T)$

NME : Nombre de Mots composant l'Extrait (taille de l'extrait)

$P(W) = \prod_{W_i \in E^T} P(W_i \notin contexte) = \prod_{W_i \in E^T} (ND - ND(W_i)/ND)^r$

Lors de l'évaluation de la qualité du résumé construit conformément à la politique intermédiaire, la performance de ce dernier est égale à la somme, sur tous les extraits comportant un pourcentage P ou plus des classes insérées dans le résumé, des produits des probabilités des classes constituant chaque extrait.

6.6 Les expériences

Maintenant que nous avons exposé notre idée d'introduction du contexte dans notre méthode de construction de résumés textuels, nous réalisons un certain nombre d'expériences. Ces dernières nous permettent de construire des résumés appropriés aux différents politiques, et d'évaluer leurs performances par rapport à notre PRM.

An – Prix – Servie – Trente-sept – Consommation – Déverser – Construit – Nitrate
Plombier – Désinfectée – Robinet – Arrondissement – Potable – Bactérie – Paysan – Fuit
Stoppe – Enterrer – Infecté – Fortune

Figure 6.4: Exemple d'un résumé textuel construit selon la politique souple avec des extraits comportant quatre mots en présence d'un contexte de dix documents.

La Figure 6.4 un exemple de résumé construit avec des extraits de quatre mots en présence d'un contexte composé de dix documents ($r = 10$) sélectionnés aléatoirement du Web. La méthode de construction est basée sur une mesure de couverture conditionnelle pondérée. A chaque étape de sélection nous insérons au résumé le mot représentant de la classe ayant la plus grande couverture conditionnelle dans le texte et la moins présente dans le contexte. Cette combinaison

de couverture avec la probabilité de non-appartenance de la classe au contexte nous permet de choisir à chaque étape du procédé de type «greedy» un optimum local. Ce qui nous permet donc d'obtenir à la fin de la construction un résumé sous-optimal ayant de bonnes performances suivant la politique souple.

An – Pneumonie – Docteur – Agréable – Foie – Mauvaise – Passion – Défend – Boby
Considérée – Prix – Vieux – Business – Adieu – Retraite – Gelée – Sens
Trente-sept – Comptent – Trottoirs

Figure 6.5: Exemple d'un résumé textuel construit selon la politique souple avec des extraits comportant quatre mots en présence d'un contexte de dix documents.

Le deuxième résumé présenté dans la figure 6.5 est aussi construit avec des extraits de quatre mots en utilisant un contexte de 10 documents. Cependant lors de la construction la mesure utilisée est différente de la couverture conditionnelle. C'est une mesure plus convenable aux politiques stricte et intermédiaire. Cette mesure consiste en la combinaison de la mesure utilisée dans le cas de la politique intermédiaire sans contexte avec la probabilité que les mots composant un extrait donné ne soient pas similaires à des mots appartenant au contexte.

Longueur de l'extrait	Politique Souple d_1^T	Politique Intermédiaire $d_3^T \quad n >= 50\%$	Politique Intermédiaire $d_3^T \quad n >= 75\%$	Politique Stricte d_2^T
1	11.41%	11.41%	11.41%	11.41%
2	12.88%	11.91%	2.72%	2.72%
3	11.16%	3.23%	0.85%	0.85%
4	8.99%	2.81%	0.97%	0.55%
5	6.80%	0.80%	0.60%	0.50%
6	4.98%	0.60%	0.40%	0.30%
8	2.38%	0.24%	0.19%	0.14%

Tableau 6.2: Performance de résumé (construit et évalué avec contexte)

Le tableau 6.2 donne les résultats d'évaluation des résumés (construits avec notre méthode de PRM combiné avec l'utilisation du contexte en utilisant deux

mesures différentes) en utilisant différentes règles de décision suivant la politique considérée. Les règles de décision utilisées dans cette expérience lors de l'évaluation tiennent compte de la présence d'un contexte composé de dix documents aléatoires ($r = 10$).

A partir des résultats que nous avons enregistrés, nous avons constaté que les performances calculées des résumés quelle que soit la longueur des extraits utilisés sont faibles(entre 0.14 et 12.88). A l'exception du premier cas où la politique souple est adoptée, les résultats des deux autres politiques d'évaluation chutent rapidement vers un taux avoisinant le 1% et parfois même le 0%. L'augmentation de la taille du résumé est évidemment une façon d'améliorer les performances, mais nous ne savons pas quelle est la taille maximale raisonnable d'un résumé. Dans le cas où la taille du résumé atteindrait une centaine de mots, il n'est plus clair que le principe de reconnaissance maximale puisse être considéré comme raisonnable du point de vue de l'utilisateur.

Malgré ces premiers résultas reflétant des performances faibles en fonction de notre PRM, l'utilisation du contexte pour détecter et éliminer les mots ambigus nous paraît toujours une opération raisonnable et importante pour la distinction du contenu d'un document donné par rapport à l'ensemble des documents qui peuvent exister. Toutefois, ces résultats sont nos premiers essais pour introduire les informations textuelles dans le procédé de construction automatique de résumé. Dans le but de garder l'utilisation du contexte mais avec un nouveau angle de vue, nous avons décidé de faire une évaluation croisée. Dans cette nouvelle expérience, nous prenons en compte des résumés créés sans l'utilisation de contexte et nous les évaluons en employant un contexte composé de 10 documents. Dans ce cas de figure, les méthodes de construction seront basées uniquement sur le PRM en utilisant les couvertures conditionnelles dans le cas de la politique souple ou bien la deuxième mesure proposée dans le cas de la politique stricte ainsi qu'intermédiaire. Cependant, lors de l'évaluation, ce sont les règles probabilistes qui sont utilisées.

Dans le tableau 6.3 les résultats de cette expérience sont exposés. En comparant avec le tableau 6.2, ces résultats montrent que globalement les performances des résumés construits et évalués avec contexte sont légèrement meilleures que celles obtenues par une évaluation croisée. Donc, il est plus intéressant d'avoir

Longueur de l'extrait	Politique Souple d_1^T	Politique Intermédiaire d_3^T $n >= 50\%$	Politique Intermédiaire d_3^T $n >= 75\%$	Politique Stricte d_2^T
1	9.80%	9.80%	9.80%	9.80%
2	11.56%	10.60%	2.19%	2.19%
3	10.13%	2.72%	0.52%	0.85%
4	7.86%	2.55%	0.72%	0.21%
5	6.35%	0.65%	0.30%	0.21%
6	4.48%	0.36%	0.18%	0.15%
8	2.33%	0.11%	0.05%	0.05%

Tableau 6.3: Performance du résumé (construit sans contexte et évalué en utilisant les informations du contexte)

les mêmes contraintes lors des phases de construction et d'évaluation, car les deux sont basées sur le même principe (PRM).

De cette observation, il ressort que lors de la création, l'application de la politique souple avec contexte semble plus appropriée que la politique souple sans contexte, malgré les petites différences lors de l'évaluation. Ceci est conforme à nos attentes sachant que les mots appartenant au résumé construit avec le contexte sont spécifiques au document correspondant tandis que le résumé construit sans contexte comporte plutôt les mots les plus fréquents (communs). Il est intéressant de noter que des résultats similaires ont été obtenus lors de la construction de résumés visuels.

Par ailleurs, si nous comparons les résultats du tableau 6.3 avec ceux du tableau 6.1 où la création et l'évaluation étaient réalisées sans la contribution du contexte, nous observons les points suivants. Les performances exprimées dans le premier tableau sont nettement plus importantes que celles du tableau 6.3. Cette première constatation indique que la construction de résumés textuels sans utilisation de contexte permet aux mots communs (présents dans plusieurs documents) d'être sélectionnés et insérés dans le résumé final. Cette présence est la cause de la grande diminution des performances lors de l'évaluation avec contexte vu que ces mêmes mots ont une forte probabilité d'appartenir au contexte utilisé.

Suite à cette comparaison, nous notons aussi que dans le cas de la politique souple, les performances des résumés augmentent en fonction de la longueur des extraits utilisés dans le premier cas (création et évaluation sans contexte) tandis que c'est le contraire dans le deuxième cas (création sans contexte, évaluation avec contexte). Nous expliquons ce phénomène, par le fait que paradoxalement, l'augmentation de la longueur des extraits utilisés augmente la probabilité de la similarité d'au moins un mot de l'extrait à un mot du résumé dans le premier cas ainsi que la probabilité de la présence dans l'extrait d'un mot similaire à un autre mot ou plus appartenant au contexte dans le deuxième cas.

En résumant cette étude, nous constatons que la création et l'évaluation de résumés textuels sans utilisation de contexte donne des valeurs de performances plus élevées que celles correspondant aux cas d'utilisation de contexte ou d'évaluation croisée. Malgré ces grandes performances dans le cas de la politique souple la signification pratique de ces derniers reste inappropriée à notre application de reconnaissance maximale car l'utilisation d'un document isolé du contexte général est très contraignante.

D'autre part, si l'objectif final de notre création de résumés textuels est d'obtenir un résumé comportant des informations tel qu'il peut substituer le contenu du document original, la méthodologie qui consiste à la combinaison du PRM avec le contexte n'est pas recommandable.

Certainement, une alternative consistera à décrire le texte comme un ensemble de sujets, et pas simplement comme étant une succession de mots, ainsi la mesure sera basée sur la présence des sujets dans le résumé au lieu de la présence des mots dans ce dernier (ceci nécessite des méthodes de traitement de texte plus élaborée, telle que l'indexation sémantique ou d'autres techniques de classification, et se situe hors de nos travaux de recherche actuels).

6.7 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle approche de construction de résumés textuels basée sur le principe de reconnaissance maximale. Nous avons introduit l'idée d'utiliser le contexte pour construire des résumés textes plus discriminatoires. Notre analyse des résultats des expériences, montre que parmi

les différentes alternatives d'algorithmes (les différentes politiques proposées ainsi que l'utilisation ou non du contexte), les performances les plus intéressantes sont obtenues en utilisant la politique souple avec contexte.

Chapitre 7

Combinaison de la vidéo et du texte

Dans ce chapitre nous présentons notre méthode de construction et d'évaluation de résumés multimédia. Nous adaptons notre principe de reconnaissance maximale que nous avons déjà appliqué séparément aux deux flux vidéo et texte afin de l'appliquer conjointement au deux média cités. Ensuite nous proposons une variante de cette méthode afin de gérer l'espace d'affichage qui contiendra le résumé multimédia, en essayant de trouver le meilleur compromis entre les mots et les images. Cette combinaison des deux flux nous permet d'enrichir et d'améliorer la qualité du résumé. Aussi, elle nous permet de former un résumé audio-visuel où les segments audio sont sélectionnés en fonction des mots appartenant au résumé multimédia construit, les segments vidéos sont sélectionnés en fonction des images insérées dans le même résumé multimédia.

7.1 Reconnaissance maximale de composants

Notre méthode de création et d'évaluation de résumés vidéo-textuels est basée sur le principe de reconnaissance maximale PRM. Cette fois-ci, nous prenons en compte une grande partie de l'ensemble des composants d'un document multimédia. Ces composants sont tirés des flux vidéo et texte. Le flux texte peut correspondre à une transcription du flux audio associé au flux vidéo ou un regroupement des sous-titres (le télé texte) sous forme d'un texte. Le résumé multimédia R^M résultant est composé d'un ensemble d'éléments des deux flux,

c'est-à-dire qu'un élément peut être une image prise du flux vidéo ou un mot tiré du texte. Le résumé créé nous permet d'avoir une idée générale du contenu de la vidéo originale ainsi que des informations sémantiques en faisant une combinaison des images et mots sélectionnés. Un extrait E^M est défini comme étant un segment de taille prédéfinie tiré de la vidéo originale, ce qui représente une succession d'images ayant une relation de voisinage associée à un ensemble de mots extraits du télé texte synchronisé à ces images dans la vidéo originale. Afin d'appliquer le PRM, nous devons définir une règle de décision qui peut être utilisée pour valider si un extrait provient du document multimédia qui correspond au résumé ou non.

En premier lieu, nous proposons la règle de décision qui consiste à dire que: «L'utilisateur décide qu'un extrait E^M est pris du document original si au moins un élément de l'extrait que ce soit une image ou un mot est similaire à un ou plusieurs éléments appartenant au résumé.» Ceci suppose qu'une image et un mot ont la même importance aux yeux d'un utilisateur; il faut noter qu'une image peut présenter plus d'information qu'un simple mot en étant spécifique au document d'où elle provient. Nous commençons tout d'abord par tester cette règle de décision. La taille du résumé global est définie en fonction des besoins des utilisateurs ou en fonction des capacités du dispositif sur lequel sera affiché le résumé. A ce stade de travail, nous faisons abstraction de la nature de l'élément sélectionné sans prendre en compte la différence entre un mot ou une image . Afin d'appliquer le PRM nous proposons une expérience dont le scénario est suivant {Figure 7.1}:

- Nous présentons le résumé vidéo-textuel R^M à l'utilisateur sous forme d'un ensemble d'images et de mots.
- Ensuite, un extrait E^M, d'une durée prédéfinie d, choisi aléatoirement du document original est montré à ce dernier.
- Nous demandons à l'utilisateur de deviner si l'extrait E^M qu'il vient de voir est tiré de la même vidéo V qui correspond au résumé R^M ou non.

D'après notre règle de décision, le comportement de l'utilisateur doit être le suivant:

- Si au moins une image ou un mot de l'extrait E^M est similaire à une image

Figure 7.1: Scénario de l'expérience de reconnaissance maximale.

ou mot du résumé R^M, il peut répondre correctement (notons que sa réponse est correcte).

- Si ce n'est pas le cas, il est en doute et ne peut procurer de réponse.

  La probabilité des réponses correctes n'est autre que la performance attendue de l'utilisateur dans cette expérience. Donc la performance est le nombre d'extraits pris de la vidéo originale contenant soit une image similaire à une ou plusieurs images du résumé soit un mot appartenant au même résumé multimédia.

7.2 Création Automatique d'un résumé vidéo-textuel

Maintenant que l'expérience de reconnaissance maximale de composants est définie, nous présentons d'une manière formelle un processus de construction automatique de résumés multimédias. Rappelons que le facteur principal de la sélection des éléments composant le résumé est la règle de décision. Notre objectif est de maximiser la performance des éléments sélectionnés pour constituer le résumé $\hat{R}^M$

comme décrit précédemment.

Faisons l'hypothèse que les extraits que nous considérons ont une durée d. Si la vidéo comporte N secondes, nous avons $N-d+1$ différents extraits, nous faisons un sous-échantillonnage en prenant une image par seconde ce qui implique qu'un extrait de durée d secondes contiendra d images, le nombre de mots correspondant à un extrait de durée d est différent d'un extrait à l'autre en fonction de la longueur des mots, le silence, la musique, le bruit, etc. . . :

- E_1 contient les images $I_1, I_2, ...I_d$ et les mots associés $M_1, ...M_k$,
- E_2 contient les images $I_2, I_3, ...I_{d+1}$ et les mots associés $M_i, ...M_l$,
- Et ainsi de suite jusqu'à E_{N-d+1} qui contient les images $I_{N-d+1}, I_{N-d+2}, I_N$ et les mots synchrones à cette extrait $M_r, ...M_x$.

Nous calculons une matrice de similarité des images traitées en calculant la distance L1 entre les vecteurs caractéristiques des différentes images. Deux images sont considérées comme étant similaires si et seulement si la distance entre leurs vecteurs caractéristiques respectifs $H(I_i), H(I_j)$ est inférieure ou égale à un seuil de similarité visuelle prédéfini S_{sv}.

$$I_i \; et \; I_j \; sont \; similaires \Longleftrightarrow Dis(H(I_i), H(I_j)) \leq S_{sv} \tag{7.1}$$

En ce qui concerne les mots, deux mots sont considérés comme étant similaires si et seulement s'ils sont orthographiquement identiques. Dans une version plus évoluée, on pourrait par exemple utiliser un algorithme d'analyse morphologique pour reconnaître si deux mots sont des formes déclinées d'une même racine, et dans ce cas, les considérer comme similaires. Dans le chapitre précédent qui traite le problème de la construction de résumés vidéo en utilisant uniquement l'information textuelle, nous avons réalisé une étude morphologique du document traité. Ce dernier était en langue française. Dans ce chapitre les documents utilisés sont en langue anglaise, donc nous avons estimé que nous pouvons nous en passer d'une telle études pour des premiers travaux.

$$M_i \; et \; M_j \; sont \; similaires \Longleftrightarrow M_i = M_j \tag{7.2}$$

Pour le flux textuel, une phase de pré-traitement est réalisée, consistant à éliminer quelques mots communs en utilisant une «stop-list». Nous ignorons ces

mots communs lors de notre processus de construction parce qu'ils n'apportent aucune information sémantique spécifique au document multimédia traité.

La figure 7.2 illustre la relation entre les extraits, les images et les mots.

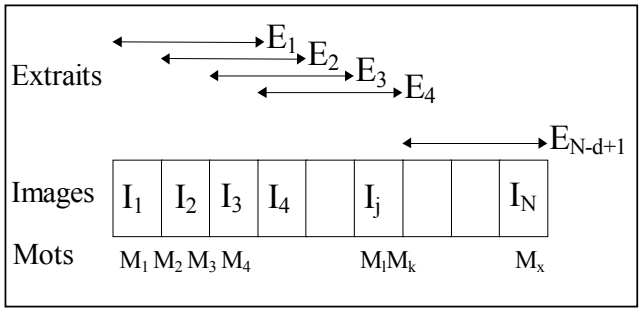

Figure 7.2: Relations entre les images, les mots et les extraits.

Nous définissons la couverture $Cov(I)$ d'une image I comme étant le nombre d'extraits, qui contiennent au moins cette image ou une image qui lui est similaire:

$$Cov(I) = Card\{i : \exists j \;\; I_j \in E_i \; et \;\; I_j \;\; similaire\ à \;\; I\} \tag{7.3}$$

Nous déterminons la couverture $Cov(M)$ d'un mot M comme étant le nombre d'extraits contenant au moins une occurrence de ce mot:

$$Cov(M) = Card\{i : \exists j \;\; M_j \in E_i \; et \;\; M_j = M\} \tag{7.4}$$

Nous désignons la couverture $Cov(El)$ d'un élément El qu'il soit une image I ou un mot M comme étant le nombre d'extraits contenant au moins cet élément El ou un élément lui est similaire:

$$Cov(El) = Card\{i : \exists j \;\; El_j \in E_i \; et \;\; El_j \;\; similaire\ à \;\; El\} \tag{7.5}$$

Nous définissons la couverture d'un ensemble d'éléments $El_1, El_2, ..., El_k$ comme étant le nombre d'extraits qui contiennent au moins un élément El_i similaire à l'un de ces éléments:

$$Cov(El_1, El_2, ...El_k) = Card\{i : \exists j \; El_j \in E_i \; et \; \exists l \in [1,k] \; El_j \; similaire\ à \; El_l\} \tag{7.6}$$

Si un résumé vidéo-textuel R^M est composé des éléments $El_1, El_2, ...El_k$, tel que El_j une image I ou un mot M, ceci mène à une performance Perf de système égale à:

$$\text{Perf} = \frac{1}{(N-d+1)} Cov(El_1, El_2, ...El_k) \tag{7.7}$$

Le résumé optimal $\hat{R}^M$ est alors simplement celui qui maximise:

$$\hat{R}^M = \underset{El_1, El_2, ...El_k}{\arg\max} \frac{1}{(N-d+1)} Cov(El_1, El_2...El_k) \tag{7.8}$$

7.2.1 Algorithme de Construction

Le résumé optimal de taille k peut être obtenu en faisant une énumération de tous les ensembles de k éléments $\{El_1, El_2, ...El_k\}$ puis en conservant le meilleur ensemble. Sachant que l'énumération de la totalité des ensembles semble très coûteuse en temps de calcul, il est plus judicieux de sélectionner minutieusement l'ordre dans lequel les éléments sont sélectionnés, ainsi les meilleures solutions sont retrouvées rapidement.

Nous rappelons que la construction de résumés optimaux dans des temps raisonnables par rapport à la disponibilité des usagers est difficile à mettre en œuvre avec les moyens actuels. Pour ces raisons nous nous contentons de la construction d'un résumé sous-optimal par rapport à notre principe de reconnaissance maximale.

Nous utilisons une heuristique analogue à celle présentée dans le chapitre 4 où seulement les informations visuelles étaient utilisées, cette fois-ci en remplaçant une image par un élément (mot/image). Cette heuristique nous permet d'obtenir si nous le désirons une solution optimale. Le déroulement de cette dernière est comme suit:

Si un élément El_m est rajouté à un ensemble existant $\{El_1, El_2, ...El_{m-1}\}$, nous pouvons définir la couverture conditionnelle comme étant la contribution de cet élément dans la couverture finale de cet ensemble:

$$Cov(El_m|El_1El_2...El_{m-1}) = Cov(El_1El_2...El_m) - Cov(El_1El_2...El_{m-1})$$
$$= Card\left\{ \begin{array}{l} i : \exists j \ El_j \in E_i \ et \ El_j \ similaire \ à \ El_m \\ et \ \ \forall El \in E_i \forall j = 1, 2, ...m-1 \ \ El \ pas \ similaire \ à \ El_j \end{array} \right\} \quad (7.9)$$

Alors, la couverture de l'ensemble des éléments $\{El_1, El_2, ...El_k\}$ peut être calculée comme suit :

$$Cov(El_1...El_k) = Cov(El_1) + Cov(El_2|El_1) + ... + Cov(El_k|El_1...El_{k-1}) \quad (7.10)$$

L'algorithme que nous proposons pour la construction du résumé optimal procède comme suit:

Etape 1: Débuter le processus avec un ensemble vide d'éléments.

Etape 2: Ordonner les éléments qui n'ont pas encore été sélectionnés en fonction de leur couverture conditionnelle décroissante par rapport à l'ensemble courant.

Etape 3: Essayer d'ajouter à tour de rôle chaque élément à l'ensemble courant. Si la taille désirée du résumé est atteinte, remplacer la solution courante par l'ensemble courant s'il possède une couverture plus importante. Sinon revenir d'une manière récursive à l'étape 2 afin de continuer l'énumération.

Etapes 4: Lorsque tous les éléments sont essayés, revenir à l'étape 2 pour continuer l'énumération.

Durant la procédure de retour, il est possible d'éviter certaines énumérations et de gagner un peu de temps de calcul en notant que la relation suivante est toujours maintenue si m<k:

$$Cov(El|El_1El_2...El_{m-1}El_m...El_k) \leq Cov(El|El_1El_2...El_{m-1}) \quad (7.11)$$

donc

$$\begin{aligned} Cov(El_1El_2...El_{m-1}El_m..El_k) &= Cov(El_k|El_1El_2..El_{k-1}) + Cov(El_{k-1}|El_1El_2..El_{k-2}) \\ &+... + Cov(El_m|El_1El_2...El_{m-1}) + Cov(El_1El_2...El_{m-1}) \\ &\leq Cov(El_k|El_1El_2...El_{m-1}) + Cov(El_{k-1}|El_1El_2...El_{m-1}) \\ &+... + Cov(El_m|El_1El_2...El_{m-1}) + Cov(El_1El_2...El_{m-1}) \end{aligned} \tag{7.12}$$

Cette inéquation entraîne une borne supérieure pour la meilleure solution qui peut être construite en étendant l'ensemble $\{El_1, El_2, ...El_{m-1}\}$. Si la couverture avec cette borne supérieure est plus petite que la couverture de la meilleure solution courante, alors l'énumération peut être arrêtée à ce niveau. Ceci diminue le taux de calcul requis, en préservant le fait que cet algorithme est optimal.

Notons que l'algorithme commence en sélectionnant l'élément El_1 ayant la couverture maximale(cet élément peut être une image ou un mot), puis El_2 qui a une couverture conditionnelle maximale par rapport à l'élément El_1 et ainsi de suite jusqu'à El_k. La première solution complète retrouvée est alors le résultat d'une série de choix avec le principe de type «greedy».

Une fois que le meilleur ensemble d'éléments est trouvé par un processus de type «greedy», nous pouvons le maintenir comme étant une solution sous-optimale. Si nous désirons obtenir une solution optimale qui maximise la couverture du résumé multimédia par rapport à la tâche de reconnaissance maximale, nous réitérons le processus de sélection en remplaçant chaque élément sélectionné par l'ensemble des éléments non sélectionnés à tour de rôle. En pratique, nous gardons la première solution de type «greedy» afin d'avoir des temps de construction de résumés multimédia raisonnables par rapport au temps d'attente des utilisateurs.

7.3 Expériences

Nous effectuons quelques expériences pour tester notre approche de construction de résumés multimédias combinant la vidéo et le texte basée sur le principe de reconnaissance maximale. Lors de ces expériences notre règle de décision consiste à décider l'appartenance d'un extrait donné à une vidéo originale si et seulement si un élément de l'extrait est similaire à un élément du résumé correspondant au

document multimédia, cet élément pouvant être une image ou tout simplement un mot. Les images sont représentées par des vecteurs caractéristiques, chaque vecteur est un histogramme de blobs de couleurs 11x11. La similarité visuelle est jugée d'une façon mathématique en utilisant la distance L1 et le même seuil de similarité que nous avons calculé lors de la construction des résumés basés uniquement sur les informations visuelles dans le chapitre 4. Ce seuil était déterminé à l'aide d'une expérience réalisée avec l'intervention de quelques personnes pour avoir un rapprochement avec le jugement humain de la similarité des images. Les documents multimédias utilisés sont:

- Un documentaire «Cooking» représenté par la vidéo A,
- Un deuxième documentaire «Predators» représenté par la vidéo B,
- Un troisième documentaire «Andes to Amazon» représenté par la vidéo C,
- Un premier film "Mission Impossible» représenté par la vidéo D,
- Un autre film «Young Americans» représenté par la vidéo E,
- Un journal télévisé de la chaîne BBC représenté par la vidéo F.

Ces documents nous ont été fournis dans le cadre du projet Européen «SPATION: Services Platforms and Applications for Transparent Information management in an in-home Network».

http://www.extra.research.philips.com/euprojects/spation/

Les durées en secondes des vidéos utilisées sont reportées dans le tableau 7.1ci-dessous:

Cooking	Predators	Andes to Amazon	Mission Impossible	Young Americans	BBC News
3713sec	1860sec	3000sec	6337sec	5812sec	1495sec

Tableau 7.1: Durées des différents documents multimédia exprimées en secondes

Pour chacun des documents vidéos cités ci-dessus, nous avons construit des résumés statiques sous forme d'une collection d'images et de mots de différentes tailles $k = 10, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100$. Aussi pour chaque résumé de taille k, nous avons testé les durées d'extraits d de $5, 10, 15, 20, 25, 30, 40$ *et* 60 secondes.

Figure 7.3: Résumé de «Andes to Amazon» ($k = 10$ & $d = 30sec$).

La figure ci-dessus 7.3 représente un exemple de résumé parmi tous ceux que nous avons construits avec notre algorithme basé sur le principe de type «greedy» ainsi que le PRM. Ce résumé correspond au documentaire «Andes to Amazon», il est composé de 10 éléments (entre images et mots). La durée des extraits utilisés lors de sa création est égale à 30 secondes. A ce niveau de l'étude, nous n'avons aucun contrôle sur la proportion entre les images ou les mots sélectionnés, nous fixons uniquement la taille globale du résumé sans aucune contrainte imposée par rapport à sa composition. A chaque étape de sélection, notre algorithme calcule les couvertures conditionnelles de chaque élément que ce soit une image ou un mot en fonction des éléments déjà choisis, ensuite il les trie et rajoute au résumé l'élément dont la couverture conditionnelle est maximale quelle que soit sa nature (image ou mot). Comme nous utilisons le principe de type «greedy» qui permet de prendre à chaque fois l'optimum local, les résumés construits de tailles différentes en utilisant la même durée d'extraits sont inclus les uns dans les autres, c'est-à-dire un résumé de taille k inclut forcément tous les résumés de taille inférieure à sa taille k. La figure 7.4 montre un résumé de taille égale à 15, créé en utilisant un extrait de durée égale à 30. Nous remarquons que ce résumé inclut entièrement le résumé composé de 10 éléments présenté dans la figure 7.3, construit en utilisant la même durée d'extrait de 30 secondes, plus 5 autres éléments (une image et quatre mots).

FOREST – MONKEYS – ANIMALS – ANTS
JUNGLE – WATER – TAKE – GROUND
MACAWS – PLAYS – EAT – MONKEY - USE

Figure 7.4: Résumé de «Andes to Amazon» ($k = 15$ & $d = 30sec$).

Par ailleurs, le nombre de mots ou d'images que comporte un résumé donné diffère d'un résumé à l'autre. La composition de chaque résumé dépend de deux paramètres qui sont la taille globale du résumé vidéo-textuel ainsi que la durée de l'extrait utilisé dans le procédé de calcul des couvertures conditionnelles. Les deux graphes représentés ci-dessous, cf. figure 7.5, représentent respectivement les pourcentages de mots et d'images dans les résumés résultants de notre algorithme de création appliqué toujours au documentaire «Andes to Amazon» en fonction de la durée des extraits utilisés lors de la sélection. Chaque tracé correspond à une taille fixe de résumé.

Nous avons constaté lors de ces expériences que parfois, pour des grands résumés ou quand les extraits utilisés sont de très grande taille, une couverture de 100% est atteinte avant que tous les éléments soient sélectionnés. Tous les extraits possibles tirés de la vidéo originale contiennent au moins un mot ou une image similaire à un des éléments composant le résumé courant. Sachant que la couverture maximale est atteinte, l'ajout d'un élément non sélectionné n'augmentera pas le nombre d'extraits reconnus grâce au résumé. Donc la couverture conditionnelle de chaque élément de l'ensemble des mots et des images non sélectionnés à ce niveau de construction est égale à 0. Nous avons décidé d'arrêter notre algorithme à ce stade car nous n'avons pas de raisons particulières et objectives par rapport à notre principe de reconnaissance maximale pour choisir un élément ou

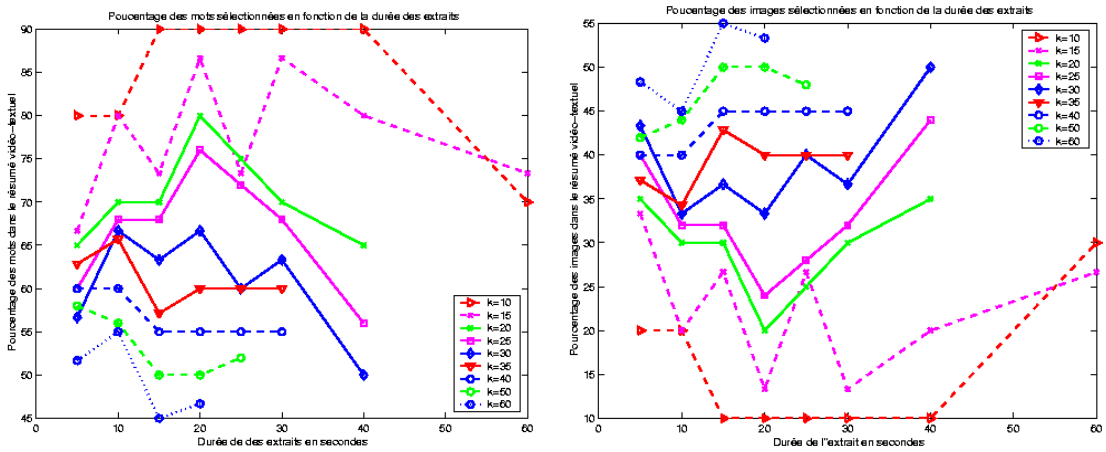

Figure 7.5: Pourcentage des mots et des images dans le résumé de «Andes to Amazon» en fonction de la durée des extraits.

un autre pour le rajouter dans le résumé actuel dont la taille est inférieure à la taille désirée.

Nous observons dans le graphe de gauche de la figure 7.5 qu'en général le pourcentage des mots inclus dans le résumé du documentaire «Andes to Amazon» diminue en fonction de la taille de ce dernier lorsque la durée des extraits utilisés dans le processus de construction est la même pour différentes tailles. Puisque les pourcentages des mots et des images sont complémentaires et comme le certifie le graphe de droite de la figure 7.5 le pourcentage des images croît en fonction de la taille globale du résumé pour la même durée des extraits utilisés lors de la création. Nous expliquons ce phénomène par le fait que notre calcul de couverture conditionnelle exclue tous les extraits comportant des éléments déjà sélectionnés conjointement avec l'élimination des mots communs que nous avons effectuée dans la phase préliminaire du traitement de la transcription textuelle.

Avant d'essayer d'analyser, d'argumenter ou de comprendre les différents résultats obtenus par notre processus de construction de résumés vidéos-textuels suivant la vidéo prise en considération parmi les six traitées nous rappelons les trois points suivants:

1. Lors de cette étude, nous combinons les informations visuelles et textuelles. Contrairement aux informations visuelles qui sont continues par rapport à

l'axe du temps, les mots ne sont pas présents partout, car certaines parties de la vidéo correspondent à de la musique, du bruit, ou tout simplement du silence. Le tableau 7.2 suivant représente le pourcentage de l'éventuelle présence ou non des mots par rapport au flux visuel.

	Andes to Amazon	Predators	Cooking	Mission Impossible	Young Americans	BBC News
Taux de mots	61,84%	60,98%	77,97%	69,86%	83,59%	95,86%
Autres	38,16%	39,02%	22,03%	30,14%	16,41%	4,14%

Tableau 7.2: Pourcentage du temps d'apparition du texte dans la vidéo

De plus, l'élimination des mots communs provoque une diminution considérable du nombre des mots conservés après cette phase. Cette diminution est estimée à la moitié du nombre global des mots de la transcription textuelle pour chacun des documents vidéos. Le tableau 7.3 indique le nombre initial des mots composant les transcriptions textuelles correspondantes aux six documents vidéos, le nombre de mots retenus après la phase d'élimination des mots communs en utilisant les «*stop-lists*» et enfin le nombre de mots distincts parmi les mots gardés.

	Andes to Amazon	Predators	Cooking	Mission Impossible	Young Americans	BBC NEWS
Totalité des mots	3180	2154	6843	4676	5717	3350
Mots retenus	1434	1028	2994	2100	2569	1612
Mots distincts	816	630	1323	1072	1060	1010

Tableau 7.3: Nombres de mots de chaque document à différentes phases de traitement

Nos extraits ont tous la même durée mesurée en secondes quel que soit leur positionnement sur l'axe du temps (début, milieu, ou fin de la séquence

vidéo). Comme nous avons effectué un sous-échantillonnage des images à une image par seconde, un extrait de durée d correspond à un nombre entier de d images. Cependant, le nombre de mots synchronisés à ces images est variable; il dépend de la position de l'extrait dans la vidéo ainsi que de la longueur de ces mots dans la transcription textuelle correspondante. Nous pouvons rencontrer des extraits ne contenant aucun mot. Parmi les situations qui engendrent le cas d'extraits que nous comptons comme des extraits vides, nous citons les extraits comportant de la musique, du bruit ou simplement du silence. Aussi nous mentionnons le cas où les mots qui correspondent à cet extrait sont tous des mots communs qui ont été éliminés préalablement. De même lorsque les extraits ne comportent que des mots fractionnés c'est-à-dire dans le cas où un mot commence avant le début ou se termine après la fin de l'extrait considéré. Cependant, si le mot est très long et est tronqué à chaque extrémité de l'extrait, nous le considérons comme un mot entier. Le tableau 7.4 représente, pour l'ensemble des vidéos de test, le nombre des extraits qui ne contiennent aucun mot plein ainsi que le pourcentage de ces extraits par rapport à totalité des extraits possibles selon la durée de ces derniers et celle de la vidéo traitée.

	Andes to Amazon	Predators	Cooking	Mission Impossible	Young Americans	BBC NEWS
Extrait	3000s	1860s	3713s	6337s	5812s	1495s
5s	874-29.17%	445-23.98%	467-12.59%	3285-51.87%	2390-41.15%	58-3.89%
10s	480-16.05%	219-11.83%	334- 9,02%	2126-33.60%	1530-26.37%	33-2.22%
15s	342-11.45%	158-8.60%	308-8.33%	1570-24.83%	1130-19.49%	22-1.48%
20s	256-8.59%	120-6.52%	287-7.77%	1203-19.04%	898-15.50%	12-0.81%
25s	208-6.99%	94-5.12%	267-7.24%	913-14.46%	721-12.46%	2-0.14%
30s	182-6.16%	74-4.04%	252-6.84%	693-10.99%	586-10.13%	0-0%
40s	148-5.00%	64-3.51%	222-6.04%	413-6.56%	408-7.07%	0-0%
60s	104-3.54%	44-2.43%	162-4.43%	147-2.34%	200-3.48%	0-0%

Tableau 7.4: Nombre et pourcentage des extraits ne comportant aucun mot plein selon la durée des extraits utilisée et la vidéo traitée.

2. Lors de ces expériences nous avons utilisé le même seuil de similarité pour les différentes vidéos. Ce seuil était repris des expériences de construction de résumés uniquement visuels. Malgré que ce seuil était déterminé à l'aide d'une intervention humaine lors d'une expérience basée sur la comparaison de quelques centaines de paires d'images prises de certaines vidéos qui sont différentes de celles en cours (épisodes de la série télévisée «Friends»), ce seuil n'est pas vraiment spécifique à chacune des vidéos traitées. Nous pouvons envisager d'avoir des seuils distinctifs, appropriés chacun à la vidéo correspondante selon la nature des images qui composent cette dernière, la prise de vue de ces images, les conditions d'éclairage, les dispositifs d'acquisition ainsi que d'autres paramètres. Disposer d'un seuil particulier et spécifique à chaque vidéo nous permettra certainement d'améliorer la qualité de jugement de la similarité des images appartenant à cette vidéo. Malheureusement la mise en oeuvre d'un tel procédé avec l'intervention d'utilisateurs réels, comme nous l'avons effectué précédemment, est très difficile et une estimation automatique du seuil adapté à chaque document multimédia selon les données visuelles traitées demande un travail de recherche laborieux et un temps considérable, ce qui n'est pas l'objectif premier de notre travail de recherche lors de cette thèse.

3. Selon notre principe de reconnaissance maximale, l'insertion de chaque élément sélectionné dans le résumé en cours de construction entraîne implicitement la reconnaissance de tous les extraits comportant cet élément ou un autre élément qui lui est similaire lors de l'évaluation. Dans le but d'optimiser le nombre d'extraits reconnus à l'aide du résumé final, nous rajoutons à chaque étape de sélection un élément qui additionne un nombre maximum de nouveaux extraits reconnus grâce à cet élément à l'ensemble des extraits reconnus par au moins un élément déjà sélectionné. Les deux groupes d'anciens et de nouveaux extraits sont complètement disjoints. C'est-à-dire que lors de l'élection du nouvel élément qui sera incorporé dans le résumé intermédiaire, nous négligeons totalement les extraits qui comportent au moins un élément déjà élu. Les couvertures conditionnelles des éléments non sélectionnés jusque là sont calculées avec exclusivement le reste des extraits n'incluant aucun élément du résumé courant. Ensuite

un classement des couvertures conditionnelles est effectué et le meilleur élément dont la couverture conditionnelle est la plus grande est rajouté au résumé actuel. Ainsi de suite jusqu'à ce que la taille désirée soit atteinte ou que la couverture globale du résumé soit égale à 100%.

Andes to Amazon	Predators	Cooking	Mission Impossible	Young Americans	BBC News
Forest	*1050*	*63375*	*57425*	*23750*	People
Monkeys	*15925*	Just	*144900*	*19075*	Police
Animals	*38125*	*18300*	*102375*	Get	*21025*
Ants	Prey	Good	*8675*	Right	Today
Jungle	*15750*	Get	*39900*	*54300*	Arms
Water	*1250*	*41875*	*63425*	Yeah	May
Take	*2125*	Sea	*133725*	Harris	Back
10125	Fish	Right	*64700*	Chris	Said
Ground	Long	Go	Get	*26400*	*27750*
Macaws	*5100*	*44350*	Ethan	*47225*	*29400*
Plays	People	Rosie	*35325*	*136400*	Just
Eat	*21600*	*12150*	*130175*	See	*1375*
Monkey	Young	*47950*	*1950*	*74475*	Package
Use	*1625*	*45400*	*23800*	*81850*	Thought
64950	Schoals	Way	Job	*87800*	Man

Tableau 7.5: Résumés des six documents traités(durée des extraits = 30s)
(Les numéros en italique sont les numéros des images sélectionnées)

Lors de la création des différents résumés des six vidéos traitées en utilisant diverses durées des extraits, nous avons observé l'ordre dans lequel le système sélectionne les éléments composant le résumé final. Le tableau 7.5 ci-dessus représente pour les six documents (les 3 documentaires, les 2 films et le JT) les éléments que comportent leurs résumés respectifs de taille égale à 15 créés avec des extraits de durée égale à 30 secondes affichés dans l'ordre de leur sélection. Les images sont représentées dans ce tableau par leur numéro d'ordonnancement dans le flux vidéo.

Nous remarquons que pour la première vidéo correspondant au documentaire «Andes to Amazon», le système a tendance à sélectionner en premier lieu les mots. Incontestablement, si les mots étaient sélectionnés c'est parce qu'ils ont acquis des couvertures conditionnelles supérieures à celles des images donc ils étaient classés au premier rang. Cet avantage des mots par rapport aux images lors de la construction du résumé de cette vidéo peut être dû au deuxième point mentionné préalablement qui consiste en le fait que le seuil de similarité utilisé n'est pas vraiment adapté à cette vidéo. Aussi d'après le tableau 7.4, lorsque la durée des extraits utilisés est égale à 30 secondes, seulement 6.16% parmi les extraits possibles de cette vidéo «Andes to Amazon» sont considérés comme vides. 93.4% de ces extraits contiennent donc des mots pleins et contribuent au calcul des couvertures conditionnelles des différents éléments (mots/images). Nous rappelons que ce sont des observations qui nous aident à expliquer le phénomène observé sur les graphes mais n'impliquent en aucun cas l'association d'une démarche prédéterminée du système lors de la construction du résumé final. Nous apercevons par exemple, que pour la création du résumé du film «Mission Impossible», contrairement au cas du documentaire, le système sélectionne surtout les images en premier lieu, parce qu'elles ont sûrement des couvertures plus importantes que celles des mots traités. Comme l'indique le tableau ci-dessous, c'est en fonction du document vidéo-textuel traité et de la nature des données (mots, images) que l'ordre des couvertures conditionnelles se décide.

La figure 7.6 montre les performances des résumés de différentes tailles construits en utilisant diverses durées des extraits. Ces résumés correspondent au documentaire «Andes to Amazon». Nous constatons que pour une durée d'extrait fixe, la performance du résumé augmente en fonction de sa taille. Ce qui répond à nos attentes car nous le rappelons encore une fois, notre méthode de sélection d'éléments est basée sur le principe de reconnaissance maximale ainsi que sur le principe de type «greedy» où chaque élément rajouté pour incrémenter la taille du résumé permet d'accroître la performance. Nous observons aussi que pour une taille fixe de résumé, la performance croit en fonction de la durée des extraits utilisés lors du calcul des couvertures. Nous expliquons ce phénomène par le fait que l'augmentation de la durée des extraits provoque l'augmentation

du nombre des images et des mots considérés par extrait. Donc, il est plus probable qu'un extrait contiendra au moins une image déjà sélectionnée. Chaque fois qu'un extrait contient un élément similaire à au moins un élément appartenant au résumé courant, cet extrait est considéré comme étant reconnu et contribue au calcul de la performance du résumé final.

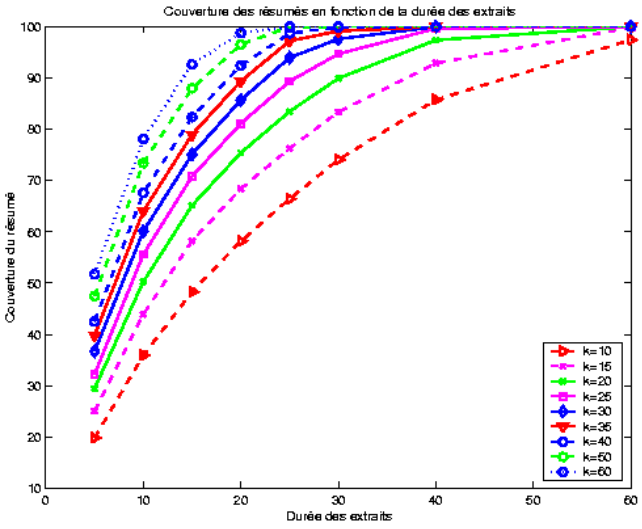

Figure 7.6: Couvertures des résumés de la vidéo «Andes to Amazon» en fonction des durées des extraits utilisés.

7.4 Gestion de l'espace d'affichage

Dans les expériences que nous avons effectuées jusque là, les images et les mots étaient considérés à égalité sans aucune différence ou préférence lors de la sélection. Nous ne pouvons apporter ni un jugement objectif ni une quantification réaliste de l'importance des images par rapport aux mots ou celle des mots par rapport aux images d'une manière générale. Par contre, il se manifeste clairement que l'espace d'affichage occupé par une image en pratique est d'usage plus important que celui utilisé par un mot et que cette relation d'espace peut être quantifiée en fonction des besoins et des préférences des utilisateurs ou en fonction du dispositif d'affichage par exemple un écran de PDA, un écran d'ordinateur ou un écran de télévision. Dans le cadre du projet européen «Spation» à lequel

notre équipe de recherche participe, notre tâche principale consiste à construire des résumés vidéo-textuels adaptés à des écran de PDA, voir la photo de la figure 7.7. Ayant déjà un procédé de construction de résumés vidéo-textuels, notre objectif est de trouver un consensus pour optimiser l'affichage des meilleurs éléments (mots/images) dans l'espace disponible.

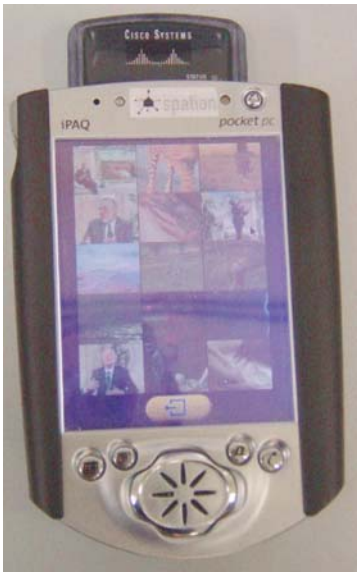

Figure 7.7: Exemple d'affichage d'un résumé multimédia sur un écran de PDA.

Dans ce but, nous proposons le changement suivant par rapport à ce que nous avons présenté préalablement: «Au lieu que la taille du résumé en construction soit estimée en un nombre fixe d'éléments, elle dépendra plutôt de la surface d'affichage».Nous considérons la surface globale qui comportera le résumé comme étant composée d'un regroupement d'un ensemble de surfaces élémentaires. Chacune de ces surfaces élémentaires correspond à l'espace que nous estimons indispensable pour l'affichage d'un mot. Nous faisons l'hypothèse que la surface nécessaire pour l'affichage d'une image est un multiple de la surface élémentaire. Dans ce cas de figure, notre algorithme de construction de résumés vidéo-textuels est très similaire au précédent avec la seule différence d'une contrainte sur la surface disponible après l'ajout de chaque élément. C'est-à-dire, le processus de construction débute par un résumé vide et une surface totale libre. Ensuite à chaque étape de sélection, il calcule les couvertures conditionnelles des

éléments (mots/images) non sélectionnés. Il les trie en fonction du rapport de leur couverture divisée par la surface disponible, puis il choisit le meilleur élément à insérer dans le résumé tel que la surface non utilisée le permet. Chaque fois qu'un élément est inséré dans le résumé, la surface libre diminue. Le taux de diminution dépend de la nature de l'élément rajouté (mot ou image), s'il s'agit d'un mot la surface totale non utilisée diminue d'une surface élémentaire, mais si c'est une image elle diminue d'un nombre multiple de la surface élémentaire. Le rapport entre la surface occupée par un mot et celle occupée par une image est un paramètre dans notre processus de création qui répond aux besoins des utilisateurs.

Afin d'étudier les résultats de cette nouvelle variante de construction de résumés vidéo-textuels où une contrainte d'espace d'affichage est imposée, nous réalisons plusieurs expériences. Pour chaque vidéo parmi les six étudiées, nous construisons divers résumés vidéo-textuels en utilisant différentes tailles d'espace global d'affichage (taille du résumé total exprimé en un nombre de surfaces élémentaires) $10, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100$; différents rapports entre la surface d'un mot et celle d'une image (la surface d'une image = $[5\ ou\ 10]*$ la surface d'un mot) ainsi que différentes durées pour les extraits utilisés lors du calcul des couvertures conditionnelles $d = 5, 10, 15, 20, 25, 30, 40, 60$.

Figure 7.8: Résumé du JT de taille égale à 40, Surface(I) = 10*Surface(M), $d = 5sec.$

La figure 7.8 représente un résumé vidéo-textuel du journal télévisé de la chaîne BBC. Ce résumé occupe quarante surfaces élémentaires, chaque image

correspond à 10 surfaces élémentaires, tandis que chaque mot occupe une seule surface élémentaire. Lors du calcul des couvertures conditionnelles, nous avons utilisé des extraits de durée égale à 5 secondes. Ceci implique la présence de 5 images par extrait suite à la phase de sous-échantillonnage que nous avons effectuée au début du processus de construction. Cependant le nombre de mots inclus dans un extrait reste variable en fonction de la position de l'extrait sur l'axe du temps. La figure 7.9 représente un autre résumé de la même séquence vidéo construit cette fois-ci en utilisant des extraits de durée égale à 15 secondes.

La première remarque que nous pouvons faire en observant les deux résumés présentés dans les figure 7.8 et 7.9 est que le nombre d'images et de mots diffère selon la durée des extraits utilisés lors du calcul des couvertures conditionnelles. Nous constatons que le résumé contient un nombre plus important de mots et moins d'images lorsque la durée des extraits utilisés a augmenté de 5sec à 10sec.

Nous avons noté, pour l'ensemble des résumés du document «JT de BBC» de taille égale à 40, que l'augmentation de la durée des extraits utilisés provoque la croissance du nombre de mots et la diminution du nombre d'images résultantes. Nous expliquons ce phénomène par le fait que la durée des extraits influe directement sur le nombre de mots pleins que contiennent tous les extraits possibles. Donc chaque fois que la durée augmente la probabilité d'obtenir des mots intégraux représentatifs, que nous considérons lors du calcul des couvertures conditionnelles des éléments composant le document multimédia (mots/images), augmente. Ceci est dû la plupart du temps à la croissance de la couverture conditionnelle de quelques mots qui se classent devant les images. Un deuxième point à signaler tient au fait que lorsqu'il reste moins de dix espaces élémentaires les images ne peuvent plus être sélectionnées. C'est-à-dire que même si à la prochaine étape de sélection, c'est une image qui représente l'optimum local en fonction de sa couverture, cette dernière ne peut être prise en compte et systématiquement remplacée par le mot ayant la plus grande couverture vis-à-vis de l'ensemble des mots non sélectionnés encore à ce niveau de construction. Le même phénomène (augmentation du nombre de mots et diminution du nombre d'images en fonction de la durée des extraits) était enregistré pour la plupart des résumés construits pour l'ensemble des vidéos traitées.

Figure 7.9: Résumé du JT de taille égale à 40, Surface(I) = 10*Surface(M), $d = 10sec$.

7.5 Gestion de la composition du résumé

Le fait d'imposer une contrainte sur la surface occupée par les images et les mots pénalise la sélection des images une fois que la surface disponible est insuffisante pour supporter cette dernière. A part cette limite, cette contrainte n'influe pas directement sur le nombre des images et le nombre de mots obtenus à la fin de la construction du résumé. Tant que l'espace est disponible, le système sélectionne l'élément ayant la plus grande couverture sans aucune autre condition. Cet élément peut être soit une image soit un mot, ceci ne dépendra que de la composition et la structuration originale du document multimédia traité. Notre objectif à ce niveau de travail est de permettre à l'utilisateur de décider de la constitution finale du résumé que nous lui construisons, en d'autres termes, lui donner la possibilité de choisir un nombre fixe d'images et respectivement un nombre fixe de mots, qui seront combinés tous ensemble pour composer le résumé de taille globale désirée. Nous avons proposé pour réaliser ce but, une nouvelle variante de notre algorithme de création de résumés vidéo-textuels. Lors du processus de la création du résumé composé de N_I images et N_M mots et à chaque étape de sélection, nous vérifions le nombre de mots et celui des images sélectionnés jusque là. Si le nombre désiré est atteint dans un cas ou dans l'autre,

nous négligeons cette classe d'éléments et nous ne la prenons plus en compte lors des prochaines sélections, c'est-à-dire, si nous avons déjà inséré dans le résumé en cours de construction N_M mots, lors de nos prochaines sélections nous gardons le meilleur élément à condition qu'il soit une image et ainsi de suite jusqu'à ce que nous complétions notre résumé, et vice-versa si le nombre total des images N_I est atteint en premier.

Dans le cadre du projet européen «Spation», nous souhaitons construire des résumés vidéo-textuels adaptés à des écrans de PDA. Nous considérons qu'un écran PDA comporte 15 emplacements consacrés à l'affichage des différents éléments composant le résumé obtenu. Chaque emplacement correspond ou bien à une image ou bien à cinq mots. Ces emplacements sont organisés sous forme d'une matrice $5 * 3$. Chaque ligne peut contenir au maximum 3 images ou 15 mots et chaque colonne supporte au plus 5 images ou 25 mots. Pour des raisons de représentation graphique et de design, nous avons décidé de construire des résumés vidéo-textuels où le nombre d'images est un multiple de 3 afin d'éviter le cas où une ligne de la matrice comportera simultanément des images et des mots. Nous mettons en place les images sélectionnées, ensuite nous complétons les emplacements libres par l'ensemble des mots choisis où chaque emplacement comprendra 5 mots. Pour chacune des six vidéos, nous construisons tous les résumés possibles en respectant cette contrainte de design. Les combinaisons possibles des résumés correspondant à chaque vidéo sont {(15images, 0mots), (12images, 15mots), (9images, 30mots), (6images, 45mots), (3images, 60mot), (0image, 75mots)}. Pour chaque combinaison, nous avons construit plusieurs résumés en utilisant diverses durées des extraits utilisés lors du calcul des couvertures conditionnelles des éléments composant le document multimédia pris en considération.

La figure ci-dessus représente un exemple de résumé vidéo-textuel créé pour le documentaire «Cooking» en utilisant des extraits de durée égale à 10 secondes. Ce dernier est composé de 9 images et 30 mots. Afin de détecter la meilleure combinaison possible pour chacun des documents multimédia pris en considération, nous comparons les performances des résumés construits pour le même document avec diverses constitutions de mots et d'images. Les combinaisons qui donnent les meilleures performances pour les six vidéos traitées ainsi que les performances

Figure 7.10: Résumé du documentaire «Cooking» composé de 9 images et 30 mots.

correspondantes sont reportées dont le tableau 7.6. Ces performances sont calculées en utilisant des extraits de durée égale à 10 secondes.

	Andes to Amazon	Predators	Cooking	Mission Impossible	Young Americans	BBC News
M_C	3I/60M	9I/30M	3I/60M	9I/30M	9I/30M	3I/60M
Perf	72.88%	94.70%	89.28%	64.23%	72.84%	98.05%

Tableau 7.6: Les meilleures combinaisons ainsi que les performances correspondantes pour les différentes vidéos (d=10sec).

D'après ce tableau, nous remarquons que pour les six vidéos étudiés, les combinaisons qui donnent les meilleures performances sont 3 images et 60 mots ainsi que 9 images et 30 mots. Pour les deux films «Mission Impossible» et «Young Americans» ainsi que le documentaire «Predators» nous obtenons un résumé où 9 emplacements sont occupés par des images contre 6 emplacements pour les mots. Pour les deux autres documentaires et le journal télévisé, les mots occupent

12 emplacements contre 3 pour les images. Notons que ces résultats sont spécifiques et dépendent de la durée des extraits utilisés lors du calcul des couvertures conditionnelles dans la phase de sélection. Le tableau ci dessous 7.7 représente le même contenu que le tableau précédent mais cette fois-ci en utilisant des extraits de durée égale à 20 secondes. Nous remarquons que pour les deux films c'est toujours la combinaison de 9 images et 30 mots qui donne les meilleures performances. Cependant pour les autres vidéos, les combinaisons ayant les meilleures performances ne sont plus les mêmes. De ce fait, nous ne pouvons assigner à chaque vidéo une combinaison qui lui soit vraiment adéquate, c'est-à-dire celle qui permet d'avoir la meilleure performance du résumé quelle que soit la durée d'extraits utilisée. En comparant les différentes combinaisons ainsi que les performances associées en utilisant différentes tailles des extraits, nous avons noté que la combinaison 9 images et 30 mots permet d'avoir souvent des résumés ayant de très bonnes performances même si ce n'est pas toujours la meilleure. Nous suggérons donc de garder cette représentation pour la construction de résumés combinés dédiés au PDA

	Andes to Amazon	Predators	Cooking	Mission Impossible	Young Americans	BBC News
M_C	6I/45M	12I/15M	12I/15M	9I/30M	9I/30M	9I/30M
Perf	92.49%	99.56%	97.64%	87.16%	89.90%	100%

Tableau 7.7: Les meilleures combinaisons ainsi que les performances correspondantes pour les différentes vidéos (d=20s).

Le schéma représenté dans la figure 7.11 résume les performances des différents résumés créés en fonction de la combinaison Mots/Images considérée. Ces performances sont calculées avec des extraits de durée égale à 20 secondes. Nous observons sur le graphe que les performances les moins bonnes sont celles qui correspondent aux résumés composés uniquement de 15 images sans aucune participation de mots. Ceci montre que par rapport à notre règle de décision la combinaison des deux médias améliore les performances des résumés créés. Nous remarquons que pour les différentes combinaisons possibles les valeurs des performances sont assez proches et il n'y a pas de grand écart.

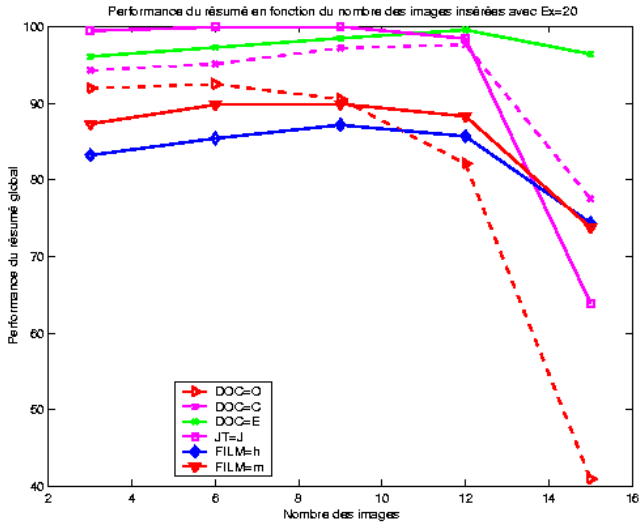

Figure 7.11: Performances des résumés en fonction de leurs constitutions ($d = 20sec$).

7.6 Amélioration du contenu par le contexte

Dans l'ensemble des résumés vidéo-textuels construits antérieurement, les mots étaient insérés dans le résumé global d'une manière individuelle sans aucune considération de leur voisinage dans le document original. Cette façon de faire permet à l'utilisateur d'avoir une idée générale mais pas précise du contenu de la vidéo traitée. Dans le but d'améliorer et de faciliter la compréhension du thème ou du sujet traité dans le document multimédia ainsi que de connaître les contextes dans lesquels étaient utilisés les mots sélectionnés afin de leur donner plus de sens, nous avons décidé en premier lieu de rajouter pour chaque mot sélectionné les deux éléments (mots/ponctuations) qui l'encadrent dans la transcription d'origine. Un mot sélectionné peut avoir plusieurs occurrences dans le même document, ce qui implique différents voisinages. Le choix du voisinage qui sera pris en compte afin d'être inséré dans le résumé parmi tout ceux possibles se fait de la manière suivante: une fois que toutes les combinaisons possibles des voisins directs sont énumérées, nous calculons le nombre d'occurrences de chaque combinaison, puis nous les trions en fonction de leur nombre d'occurrences. La combinaison ayant

le plus grand nombre d'occurrences c'est-à-dire la plus fréquente dans le document est prise en considération pour être rajoutée au mot qu'elle encadre dans le résumé final. La figure ci-dessous 7.12 représente un exemple de ce nouveau type de résumé. C'est le résumé de la même vidéo «Cooking» pour laquelle le résumé initial présenté sans ajout d'information contextuelles est présenté dans la figure 7.11, il est composé d'une combinaison de 9 images et de 30 mots encadrés par leurs successeurs et prédécesseurs dans la transcription du télé texte associée à cette vidéo. Ce résumé est construit en utilisant des extraits de taille égale à 10 secondes.

Figure 7.12: Résumé combiné du documentaire «Cooking» avec l'ajout des voisins immédiats des mots sélectionnés.

En prenant les voisins immédiats qui encadrent directement les mots sélectionnés, nous avons noté que la plupart des voisinages gardés contiennent des mots outils très fréquents qui vraisemblablement ne rajoutent pas d'informations sémantiques par rapport à celles déjà exprimées par les mots sélectionnés et insérés séparément. Par conséquent, nous avons décidé d'identifier les voisins les

plus proches qui ne sont pas des mots outils au lieu des voisins immédiats. Pour ceci, nous construisons un nouveau document, dans lequel tous les mots outils sont éliminés, nous gardons ainsi uniquement les mots que nous considérons utiles pour rajouter de l'information sémantique. Les nouveaux voisins immédiats des mots insérés dans le résumé sont certainement des mots non communs. Ensuite, pour sélectionner la meilleure combinaison des voisins parmi toutes les combinaisons possibles, nous procédons de la même manière que le cas précédent lorsque les mots outils étaient pris en considération. Nous calculons le nombre d'occurrences de chaque combinaison possible (prédécesseur + successeur) et nous gardons la plus fréquente. Enfin, nous recherchons cette combinaison dans le document original pour prendre en compte tous les mots que cette combinaison de voisins proches encadrent dont le mot sélectionné par notre algorithme basé sur la reconnaissance maximale et les insérer tous dans le résumé final. Il faut noter que pour la même combinaison de voisins proches non communs ayant la plus grande fréquence, le contenu intermédiaire peut être différent. Afin de ne pas encombrer le résumé final, nous avons décidé dans ce cas de figure de choisir la portion de texte la plus courte. Nous attirons aussi l'attention sur le fait suivant: lors de notre processus de construction automatique de résumés vidéo-textuels, toute la ponctuation était éliminée après la phase préliminaire car cette dernière n'a pas de sens sémantique et n'apporte pas d'information aux utilisateurs si elle fait partie du résumé construit. Cependant, en calculant les voisins proches non communs nous prenons en compte les ponctuations qui démarquent la fin d'une phrase comme par exemple le point, le point d'interrogation et le point d'exclamation dans le but d'éviter d'avoir des morceaux de phrases très longs ou des morceaux réunissant deux parties appartenant à de deux phrases différentes. La figure suivante 7.13 représente le même résumé que celui des figures 7.11 et 7.12 en prenant les voisins bilatéraux des mots sélectionnés délimités par deux éléments non communs (mots/ponctuations). Nous observons que les portions de phrases ou même les phrases courtes insérées dans le résumés globales à la place des mots sélectionnés rajoutent de l'information sémantique et accroissent la compréhension du thème principal du documentaire.

Figure 7.13: Résumé combiné du documentaire «Cooking» avec l'ajout d'éléments contextuels aux mots sélectionnés.

7.7 Conclusion

Dans ce chapitre, nous avons proposé une première méthode de construction de résumés vidéo-textuels en combinant l'information visuelle et textuelle. Cette méthode est toujours basée sur notre principe de reconnaissance maximale où les mots et les images sont considérés à égalité sans aucune préférence. Ensuite nous avons proposé des variantes de cet algorithme pour en premier lieu gérer l'espace d'affichage disponible, ensuite fournir à l'utilisateur un moyen de choisir la constitution du résumé désiré en désignant le nombre de mots et d'images à inclure dedans. Afin d'avoir des informations contextuelles, nous avons remplacé chaque mot du résumé par un segment d'une phrase le contenant.

Chapitre 8

Conclusion et Perspectives

Nous concluons ce rapport de thèse par un résumé global de son contenu ainsi que quelques perspectives pour des travaux futurs.

8.1 Résumé

A travers cette thèse, nous avons présenté une nouvelle méthode de création et d'évaluation automatique des résumés multimédia. Cette méthode permet de construire des résumés vidéos qui vont enregistrer des performances optimales ou quasi-optimales lors d'une évaluation extrinsèque basée sur une tâche prédéfinie. Cette évaluation est simulée par notre système en remplaçant les utilisateurs réels (qui sont censés accomplir cette tâche) par une méthode automatique grâce à un ensemble d'hypothèses faites sur leur comportement. Nous avons défini la tâche d'évaluation comme une tâche d'identification et de reconnaissance. Ensuite nous avons énoncé notre principe de reconnaissance maximale (PRM) pour la construction de résumés basée sur la tâche de reconnaissance (TR). Le principe de construction de résumés est assez générique; nous l'avons adapté aux informations visuelles, aux informations textuelles, aux deux conjointement ainsi qu'au cas du multi-vidéos.

Dans le cas des résumés visuels, nous avons testé différentes mesures de similarité visuelle des images afin de définir une bonne représentation des images ainsi qu'un seuil adéquat. Cette phase de pré-traitement nous permet d'adapter notre PRM en définissant une règle de décision adaptée à ce type de média.

Une fois que le mécanisme de construction de résumés visuel était mis en œuvre, nous l'avons modifié selon différentes contraintes afin d'obtenir quatre nouvelle méthodes de construction de résumés multi-vidéos. Ces variantes ainsi que deux autres méthodes inspirées des travaux de recherche effectués par d'autres groupes étaient comparées pour définir la plus appropriée pour cette tâche de reconnaissance. Afin de valider la qualité des résumés construits par la méthode ayant les meilleures performances suite à une évaluation automatique, nous avons élaboré une évaluation réaliste avec l'intervention d'un certain nombre d'utilisateurs réels. Cette étude nous a orienté vers les fonctions de traitement d'images qui permettront dans le futur d'améliorer les performances de notre système.

Ensuite, nous avons adapté notre principe PRM pour la construction de résumés textuels. Nous avons introduit l'idée d'utiliser le contexte pour construire des résumés textes plus discriminants. Nous avons défini différentes règles de décision en fonction d'hypothèses faites sur l'authenticité d'un mot donné.

Enfin, nous avons proposé une méthode de construction de résumés vidéo-textuels en combinant l'information visuelle et textuelle. Cette méthode est aussi basée sur notre principe de reconnaissance maximale où les mots et les images sont considérés à égalité sans aucune préférence. Des variantes de l'algorithme de construction étaient ensuite présentées pour gérer l'espace d'affichage disponible, et fournir à l'utilisateur un moyen de choisir la constitution du résumé désiré en désignant le nombre de mots et d'images à y inclure. Afin d'augmenter la compréhension du contenu de la vidéo à travers le résumé, des informations contextuelles ont été rajoutées en remplaçant chaque mot du résumé par un segment d'une phrase le contenant.

8.2 Perspectives

Nous nous sommes concentrés dans cette thèse sur l'utilisation de deux types de média: vidéo et texte. La suite de ce travail va consister à introduire le troisième type de média (l'audio). Il sera intéressant d'adapter le principe de reconnaissance maximale à l'audio dans le but de construire des résumés audio, et de combiner ensuite les informations, visuelles, textuelles et audio pour construire un résumé

multimédia optimal par rapport à la tâche d'identification d'un segment audio-visuel comportant du télé-texte. Dans ce cas de figure, le résumé construit sera présenté sous une forme dynamique (version réduite de la vidéo originale).

Il faut noter aussi que tous renforcement ou perfectionnement de la mesure de détermination automatique de la similarité des images ne peut qu'améliorer notre système de construction de résumés vidéos. Cette consolidation de la mesure de similarité peut être, par exemple, établie par l'utilisation de différents descripteurs que ceux relatifs à la couleur comme les descripteurs de textures, et les descripteurs de formes, la reconnaissance des différents objets, l'usage de descripteurs de haut niveau, et la recherche d'une ddistance plus appropriées pour le jugement de similarité.

Références bibliographiques

[AGD92] A.H.Morris, G.M.Kasper, and D.A.Adams. The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, III:17–35, 1992.

[AM99] A.Merlino and M.Maybury. An empirical study of the optimal presentation of multimedia summaries of broadcast news. *In Advances in Automatic Text Summarization, Eds. Mani, I. and Maybury, M.T. The MIT Press, 1999, pp. 391-401.*, 1999.

[BE97] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. *In proceedings of ACL'97 Workshop on intelligent, scalable text summarisation*, pages 10–17, 1997.

[BGG99] Patrick Bouthemy, Marc Gelgon, and Fabrice Ganansia. A unified approach to shot change detection and camera motion characterization. *IEEE Transaction on Ciruits and Systems for Video Technology*, VIIII:1030–1044, October 1999.

[BK97] B. Boguraev and C. Kennedy. Salience-based content characterisation of text documents. *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization*, pages 2–9, 1997.

[BMR95] R. Brandow, K. Mitze, and L. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31:675–686, 1995.

[Bou00] Nozha Boujemaa. On competitive unsupervised. *15th Ithernational Conference on Pattern Recognition*, 1:631–634, 3-7 September 2000.

[BR96] J. S. Boreczky and L. A. Rowe. Comparison of video shot boundary detection techniques. *SPIE on Storge and Retrieval from Image and Video Databases*, IV:170–179, 1996.

[cAGLO99] c. Aone, J. Gorlinsky, B. Larsen, and M.E. Okurowski. A trainable summarizer with knowledge acquired from robust nlp techniques. *In Advances in Automatic Text Summarization, Eds. Mani, I. and Maybury, M.T. The MIT Press, 1999, pp. 71-80.*, 1999.

[CGP+00] Patrick Chiu, Andreas Girgensohn, Wolf Polak, Eleanor Rieffel, and Lynn Wilcox. A genetic algorithm for video segmentation and summarization. *IEEE International Conference on Multimedia and Expo. ICME2000*, III:1329–1332, 2000.

[CI02] Janko Calic and Ebroul Izquierdo. Efficient key-frame extraction and video analysis. *IEEE International Conference on Information Technology: Coding and Computing (ITCC'02)*, pages 28–33, 8-10 April 2002.

[CN00] Patrizio Campisi and Alessandro Neri. Synthetic summaries of video sequences using a multiresolution based key frame selection technique in a perceptually uniform color space. *International Conference on Image Processing*, II:299–302, 2000.

[CSI+02] J. Calic, S. Sav, E. Izquierdo, S. Marlow, N. Murphy, and N.E. O'Connor. Temporal video segmentation for real-time key frame extraction. *IEEE International Conference on Acoustic, Speech, and Signal Processing. (ICASSP'02)*, IV:3632–3635, 13-17 May 2002.

[CSL99] Hyun Sung Chang, Sanghoon Sull, and Sang Uk Lee. Efficient video indexing scheme for content-based retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(8):1269–1279, December 1999.

[DBVF98] J.S. De Bonet, P. Viola, and J.W. Fisher. Flexible histograms: A multiresolution target discrimination model. *Procedings of SPIE*, 3370, 1998.

[DDAK99] N. D. Doulamis, A. D. Doulamis, A. D. Avrithis, and S. D. Kollias. A stochastic framework for optimal key frame extraction from mpeg video databases. *Computer Vision and Image Understanding. Special issue on content-based access for image and*

video libraries, 75:3–24, July/August 1999.

[DDK00] Anastasios D. Doulamis, Nikolaos D. Doulamis, and Stefanos D. Kollias. Efficient video summarization based on a fuzzy video content representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 10:501–517, June 2000.

[Dje02] Chabane Djeraba. Content-based multimedia indexing and retrieval. *IEEE. Multimedia*, VIIII:18–22, April-June 2002.

[Dje03] Chabane Djeraba. Association and content-based retrieval. *IEEE Transaction on Knowledge and Data Engineering*, 15:118–135, Jan-Feb 2003.

[DKAM96] Howard D.Wactlar, Takeo Kanade, Michael A., and Scott M.Stevens. Intelligent access to digital video: Informedia project. *IEEE Computer*, 29:46–52, May 1996.

[DKD98] D. DeMenthon, V. Kobla, and D. Doermann. Video summarization by curve simplification. *ACM International Conference on Multimedia*, pages 211–218, August 1998.

[DMT97] W. Ding, G. Marchionini, and T. Tse. Previewing video data: browsing key frames at high rates using a video slide show interface. *International Symposium on Research, Development and Practice in Digital Lobrairies(ISDL'97)*, pages 425–426, 1997.

[DPD98] D. Diklic, D. Petkovic, and R. Danielson. Automatic extraction of representative keyframes based on scene content. *Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems Computers*, I:877 –881, 1998.

[Duf00] Frederic Dufaux. Key frame selection to represent a video. *International Conference on Image Processing*, II:275–278, 2000.

[Edm69] H.P. Edmundson. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16:264–285, 1969.

[FT97] A. Mufit Ferman and A. Murat Tekalp. Multiscale content extraction and representation for video indexing. *SPIE on Multimedia Storage and Archiving Systems II*, 3229:23–31, 1997.

[FT99] A. Mufit Ferman and A. Murat Tekalp. Probabilistic analysis and

extraction of video content. *International Conference on Image Processing*, 2:91–95, 1999.

[FTM00] A. Mufit Ferman, A. Murat Tekalp, and Rajiv Mehrotra. Robust color histogram descriptors for video segment retrieval and identification. *IEEE Transaction on Image Processing*, 11:497–508, May 2000.

[GAT61] G.J.Rath, A.Resnick, and T.R.Savage. The formation of abstracts by selection of sentences. *American Documentation*, 12:139–143, 1961.

[GB99] Andreas Girgensohn and John Boreczky. Time-constrained keyframe selection technique. *IEEE International Conference on Multimedia Computing and Systems*, I:756–761, 1999.

[GFT97] B. Günsel, Y. Fu, and A.M. Tekalp. Hierarchical temporal vidéo segmentation and content characterization. *SPIE on Mulimedia Storage and Archiving Systems II*, 3229:46–56, 1997.

[GKA98] U. Gargi, R. Kasturi, and Antani. Performance characterization and comparison of video indexing algorithms. *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 559–565, June 1998.

[GL00a] Yihong Gong and Xin Liu. Generating optimal video summaries. *IEEE International Conference on Multimedia and Expo*, III:1559–1562, 30 July-2 August 2000.

[GL00b] Yihong Gong and Xin Liu. Video summarization using singular value decomposition. *IEEE International Conference on Computer Vision and Pattern Recognition*, II:174–180, 13-15 June 2000.

[GL01] Yihong Gong and Xin Liu. Video summarization with minimal visual content redundancies. *IEEE International Conference on Image Processing*, III:362–365, 7-10 October 2001.

[HJ99] Alan Hanjalic and Hong Jiang. An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. *IEEE Transactions on circuits and systems for video technology*, 9(8):1280–1288, December 1999.

[HLR+99] Qian Huang, Zhu Liu, Aaron Rosenberg, David Gibbon, and Behzad Shahraray. Automated generation of news content hierarchy by integrating audio, video, and text information. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, VI:3025–3028, 15-19 March 1999.

[How98] Nicholas R. Howe. Percentile blobs for image similarity. *IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 78–83, 21 June 1998.

[HRKM+97] J. Huang, S. Ravi Kumer, M. Mitra, W-J Zhu, and R. Zabih. Image indexing using color correlograms. *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 726–768, June 1997.

[HT97] Keesook J. Han and Ahmed H. Tewfik. Eigen-image based video segmentation and indexing. *IEEE International Conference on Image Processing*, II:538–541, 26-29 October 1997.

[HV98] D. Harman and E.M. Voorhees. The fifth text retrieval conference (trec-5). *National Institute of standards and Technology*, pages 500–238, 1998.

[HYM02] Benoit Huet, Itheri Yahiaoui, and Bernard Merialdo. Image similarity for automatic video summarization. *11th European Signal Processing Conference (EUSIPCO)*, III:353–356, 3-6 September 2002.

[IE99] I.Mani and E.Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, I:1–23, 1999.

[IL96] Giridharan Iyengar and Andrew B. Lippman. Videobook: An experiment in characterization of video. *IEEE International Conference on Image processing. ICIP'96*, 3:855–858, 16-19 September 1996.

[IL97a] Giridharan Iyengar and Andrew B. Lippman. Evolving discriminators for querying video sequences. *SPIE on Storage and retrieval from image and video databases V*, 3022:154–165, 13-14 February 1997.

[IL97b] Giridharan Iyengar and Andrew B. Lippman. Models for automatic

classification of video sequences. *SPIE on Storage and Retrieval for Image and Video Databases*, pages 216–227, 1997.

[IL98] Giridharan Iyengar and Andrew B. Lippman. Clustering images using relative entropy for efficient retrieval. *In Workshop on Very Low bitrate Video Coding*, October 1998.

[JBMK98] Shanon X. Ju, Michael J. Black, Scott Minneman, and Don Kimber. Summarization of videotaped presentations: Automatic analysis of motion and gesture. *IEEE Transactions on Circuits and Systems for Video Technology*, VIII:686–696, 1998.

[JLZ00] Hao Jiang, Tong Lin, and Hong-Jiang Zhang. Video segmentation with assistance of audio content analysis. *IEEE International Conference on Multimedia and Expo, ICME 2000*, 3:1507–1510, 30 July-2 August 2000.

[JV96] A. K. Jain and A. Vailaya. Image retrival using color and shape. *Pattern Recognition Journal*, 29:1233–1244, August 1996.

[KJK95] K.McKeown, J.Robin, and K.Kukich. Generating concise natural language summaries. *Information Processing and Management*, 31:703–733, 1995.

[KM02] Jungwon Kang and Russell M. Mersereau. An effective method for video segmentation and sub-shot charaterization. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, IV:3652–3655, 13-17 May 2002.

[KPC95] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. *In proceedings of the 18th ACM-SIGIR conference on research and development in information retrieval*, pages 68–73, 1995.

[KTIA98] Toshio Kawashima, Kouichi Tateyama, Toshimasa Iijima, and Yoshinao Aoki. Indexing of baseball telecast for content-based video retrieval. *International Conference on Image Processing*, 1:871–874, 1998.

[LCW03] Beito Li, Edward Chang, and Yi Wu. Discovery of a perceptual distance function for measuring image similarity. *ACM Communications on Multimedia Systems*, VIII:512–522, April 2003.

[Leh81] W.G. Lehnert. Plot units: A narrative summarization strategy. *Cognitive Science*, 4:293–331, 1981.

[LGA+99] V.Di Lecce, G.Dimauro, A.Guerriero, S.Impedovo, G.Pirlo, and A.Salzo. Image basic features indexing techniques for video skimming. *IEEE International Conference on Image Analysis and Processing*, pages 715–720, 27-29 September 1999.

[LMZW02] Xiaoye Lu, Yu-Fei Ma, Hong-Jiang Zhang, and Lide Wu. An integrated correlation measure for semantic video segmentation. *IEEE International Conference on Multimedia and Expo, ICME'02*, 1:57–60, 26-29 August 2002.

[LP96] F. Liu and W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):722–733, July 1996.

[LPE97] Rainer Lienhart, Silvia Pfeiffer, and Wolfgang Effelsberg. Video abstracting. *Communications of ACM*, 40:55–62, December 1997.

[LPE99] Rainer Lienhart, Silvia Pfeiffer, and Wolfgang Effelsberg. Scene determination based on video and audio features. *IEEE Conference on Multimedia Computing and Systems*, pages 07–11, 1999.

[LPsF97] Rainer Lienhart, Silvia Pfeiffer, and stephan Fischer. Automatic movie abstracting. *Technical Report TR-97-003*, July 1997.

[Luh58] H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, II:159–165, 1958.

[LWW00] J. Li, JZ. Wang, and G. Wiederhold. Irm: integrated region matching for image retrieval. *ACM International Conference onMultimedia*, pages 147–156, 2000.

[Mar99] D. Marcu. Discourse trees are good indicators of importance in text. *In Advances in Automatic Text Summarization, Eds. Mani, I. and Maybury, M.T. The MIT Press, 1999, pp. 123-136.*, 1999.

[May95] M.T. Maybury. Generating summaries from event data. *Information Processing and Management*, 31:735–751, 1995.

[ME00] Ken Masumitsu and Tomio Echigo. Video summarization using

reinforcement learning in eigenspace. *IEEE International Conference on Image Processing, ICIP*, II:267–270, 2000.

[Mer97] Bernard Merialdo. Automatic indexing of tv news. *Workshop on Image analysis for Multimedia Integrated services*, June 1997.

[MHYS02] Bernard Merialdo, Benoit Huet, Itheri Yahiaoui, and Fabrice Souvannavong. Automatic video summarization. *International Thyrrenian Workshop on Digital Communication, Advanced Methods for Multimedia Signal Processing*, 8-11 September 2002.

[MJ99] S.H. Myaeng and D.H. Jang. Development and evaluation of a statistically-based document summarization system. *In Advances in Automatic Text Summarization, Eds. Mani, I. and Maybury, M.T. The MIT Press, 1999, pp. 61-70.*, 1999.

[ML00] H. Martin and R. Lozano. Dynamic video abstract generation using an object dbms. *IEEE International Conference on Multimedia and Expo*, 3:1523–1526, 2000.

[MLZL02] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. *ACM International Conference on Multimedia*, pages 533–542, 1-6 December 2002.

[MM96] W. Y. Ma and B. S. Manjunath. Texture features and learning similarity. *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 425–430, June 1996.

[MM97] Mark T. Maybury and Andrew E. Merlino. Multimedia summaries of broadcast news. *Intelligent Information Systems*, pages 442–449, 1997.

[MM99] Inderjeet Mani and Mark T. Maybury. *Advances in Automatic Text Summarization*. The MIT Press, 1999.

[MSG97] P. Minel, Nugier S., and Piat G. How to appreciate the quality of automatic texte summarization. *Workshop on Intelligent Scalable Text Summarization ACL/EACL'97*, pages 25–30, July 1997.

[MT96] Peter J. Macer and Peter J. Thomas. Video storyboards: Summarising video sequences for indexing and searching of video databases. *IEE Colloquium on Intelligent Image Databases*, 2:1–5, 1996.

[NT99] Jeho Nam and Ahmed H. Tewfik. Video abstract of video. *IEEE 3rd Workshop on Multimedia Signal Processing*, pages 117–122, 13-15 September 1999.

[OH00] JungHwang Oh and Kien A. Hua. An efficient technique for summarizing videos using visual contents. *IEEE International Conference on Multimedia and Expo*, II:1167–1170, 2000.

[PJ93] C. Paice and P. Jones. The identification of important concepts in highly structured technical papers. *Sixteeth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 69–78, 1993.

[PKSW98] A. Pope, R. Kumar, H. Sawhney, and C. Wan. Video abstraction: summarizing video content for retrieval and visualization. *Conference Record of the Thirty-Second Asilomar Conference*, I:915–919, 1998.

[PLKE98] Silvia Pfeiffer, Rainer Lienhart, Gerald Kuhne, and Wolfgang Effelsberg. The moca project: Movie content analysis research at the university of mannheim. *GI Jahrentagung*, pages 329–338, 1998.

[PZ75] J.J. Pollock and A. Zamora. Automatic abstracting research at chemical abstracts service. *Journal of Chemical Information and Computer Sciences*, 15:226–232, 1975.

[PZ96] Greg Pass and Ramin Zabih. Histogram refinement for content-based image retrieval. *IEEE Workshop on Application of Computer Vision*, pages 96–102, December 1996.

[QvBS00] Richard J. Qian, Peter J.L. van Beek, and M. Ibrahim Sezan. Image retrieval using blob histograms. *IEEE International Conference on Multimedia and Expo*, 1:125–128, 2000.

[SB91] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, VII:11–32, 1991.

[SC97] John R. Smith and S-F Chang. Safe: A general framework for integrated spatial and feature image search. *In IEEE Workshop on Multimedia Signal Processing*, 1997.

[SC01] Hari Sundaram and Shih-Fu Chang. Constrained utility maximization for generating visual skims. *IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 124–131, 14 December 2001.

[Sha95] Behzad Shahraray. Scene change detection and content-based sampling of video sequences. *IST/SPIE on Digital Video Compression: Algorithms and Technologies*, 2419:2–13, February 1995.

[SK97a] Michael A. Smith and Takeo Kanade. Video skimming and characterization through the combination of image and language understanding. *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 775–781, 17-19 June 1997.

[SK97b] Michael A. Smith and Takeo Kanade. Video skimming and characterization through the combination of image and language understanding techniques. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 775–781, 1997.

[SM99] S.Teufel and M.Moens. Argumentative classification of extracted sentences as a first step towards flexible abstracting. *In Advances in Automatic Text Summarization, Eds. Mani, I. and Maybury, M.T. The MIT Press, 1999, pp. 155-171.*, 1999.

[SO96] M. Stricker and M. Orengo. Similarity of color images. *SPIE on Storage qnd Retrieval of Image and Video Databases*, 2420:381–392, 1996.

[SPD00] A. Stefanidis, A. Partsinevelos, and A. Doucette. Summarizing video datasets in the spatiotemporal domain. *11th International Workshop on Database and Expert Systems Applications*, pages 906–912, 4-8 September 2000.

[SSWW99] T. Strzalkowski, G. Stein, J. Wang, and B. Wise. A robust practical text summarizer. *In Advances in Automatic Text Summarization, Eds. Mani, I. and Maybury, M.T. The MIT Press, 1999, pp. 137-154.*, 1999.

[SZ99] Emile Sahouria and Avideh Zakhor. Content analysis of video using principal components. *IEEE Transactions on Circuits and Systems for Video Technology*, VIIII(8):1290–1298, December 1999.

[TAOS93] Y. Tonomura, A. Akutsu, K. Otsuji, and T. Sadakate. Videomap and videospaceicon: Tools for anatomizing video content. *ACM INTERCHI'93*, pages 131–141, 1993.

[TAT97] Y. Taniguchi, A. Akutsu, and Y. Tanomura. Panaramaexcerpts: Extracting and packing panoramas for video browsing. *ACM International Conference on Multimedia*, pages 427–436, 1997.

[TLD00] Candemir Toklu, Shih-Ping Liou, and Madirakshi Das. Videoabstract: A hybrid approach to generate semantically meaningful, video summaries. *IEEE International Conference on Multimedia and Expo, ICME*, III:1333–1336, 2000.

[TM99] T.Firmin and M.J.Chrzanowski. An evaluation of automatic text summarization systems. *In Advances in Automatic Text Summarization, Eds. Mani, I. and Maybury, M.T. The MIT Press, 1999, pp. 325-336.*, 1999.

[UF99] Shingo Uchihachi and Jonathan Foote. Summarizing video using a shot importance measure and a frame-packing algorithm. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, VI:3041–3044, 1999.

[UMY91] H. Ueda, T. Miyatake, and S. Yoshizawa. An interactive natural-motion-picture dedicated multimedia authoring system. *ACM SIGCHI 91*, pages 343–350, 1991.

[UU99] U.Hahn and U.Reimer. Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. *In Advances in Automatic Text Summarization, Eds. Mani, I. and Maybury, M.T. The MIT Press, 1999, pp. 215-232.*, 1999.

[VB00] C. Vertan and N. Boujemaa. Color texture classification by normalized color space representation. *15th Ithernational Conference on Pattern Recognition*, 3:580–583, 3-7 September 2000.

[VL98a] Nuno Vasconcelos and Andrew Lippman. Bayesian modeling of video editing and structure: Semantic features for video summarization and browsing. *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, 3:153–157, 1998.

[VL98b] Nuno Vasconcelos and Andrew Lippman. A spatiotemporal motion model for video summarization. *IEEE Computer Society Conference on Computer Vision and Pattern Recognnition*, pages 361–366, 23-25 June 1998.

[VXJ98] A. Vailaya, W. Xiong, and A. K. Jain. Query by video clip. *IEEE International Conference on Pattern Recognition*, 7:369–384, January 1998.

[YBL96] Minerva M. Yeung and Boon-Lock. Time-constrained clustering for segmentation of video into story units. *13th International Conference on Pattern Recognition*, III:375–38O, 25-29 Aug 1996.

[YL95a] B.L. Yeo and B. Liu. Rapid scene analysis on compressed video. *IEEE Transactions on circuit and systems for video technology*, 5:533–544, Decenber 1995.

[YL95b] Minerva M. Yeung and Bede Liu. Efficient matching and clustering of video shots. *IEEE International Conference on Multimedia Computing ans Systems*, pages 338–341, October 1995.

[YMH01a] Itheri Yahiaoui, Bernard Merialdo, and Benoit Huet. Automatic summarization of multi-episode videos with the simulated user principle. *Workshop on MultiMedia Signal Processing (MMSP'01)*, pages 461–466, 3-5 October 2001.

[YMH01b] Itheri Yahiaoui, Bernard Merialdo, and Benoit Huet. Generating summaries of multi-episodes video. *IEEE International Conference on Multimedia Expo (ICME2001)*, 22-25 August 2001.

[YMH01c] Itheri Yahiaoui, Bernard Merialdo, and Benoit Huet. Optimal video summaries for simulated evaluation. *European Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 397–401, 19-21 September 2001.

[YY97] Boon-Lock Yeo and Minerva M. Yeung. Analysis and synthesis for new digital video applications. *IEEE International Conference on Image Processing ICIP'97*, 1:1–4, 26-29 October 1997.

[YYWL95] Minerva M. Yeung, Boon-Lock Yeo, Wayne Wolf, and Bede Liu.

Video browsing using clustering ans scene transitions on compressed sequences. *SPIE on Multimedia Computing and Networking*, 2417, 1995.

[ZLSW95] H.J. Zhang, C.Y. Low, S.W. Smoliar, and J.H. Wu. Video parsing, retrieval and browsing: An integrated and content-based solution. *ACM International Conference on Multimedia*, pages 15–24, 1995.

[ZMM96] R. Zabih, J. Miller, and K. Mai. Video browsing using edges and motion. *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 439–446, June 1996.

[ZRHM98] Y. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. *IEEE International Conference on Image processing. ICIP'98*, I:866–870, 4-7 October 1998.

[ZZC96] D. Zhong, H.J. Zhang, and S.F. Chang. Clustering methods for video browsing and annotation. *SPIE on Storage and retrieval from image and video databases*, IV:239–246, 1996.

Publications Personelles

Journal

- EURASIP " Comparison of Multi-Episodes Video Summarization Algorithms" Itheri Yahiaoui, Bernard Merialdo et Benoit Huet, EURASIP Journal on Applied Signal Processing, Special issue on Multimedia Signal Processing, Vol 2003, No. 1, pages 48-55, January 2003.

Conférences Internationales

- ICMF2001 " Multi-Episodes Video Summaries" Benoit Huet, Itheri Yahiaoui et Bernard Merialdo, International Conference on Media Futures (ICMF), May 8-9, 2001, Florance, Italy.
- ISKO2001 " Automatic construction of multi-vidéo summaries" Itheri Yahiaoui, Bernard Merialdo et Benoit Huet, Filtrage et résumé automatique de l'information sur les réseaux (ISKO), 5-6 juillet 2001, Nanterre, France.
- ICME2001 " Generating Summaries of Multi-Episodes Video" Itheri Yahiaoui, Bernard Merialdo et Benoit Huet, International Conference on Multimedia & Expo (ICME2001), August 22-25, 2001, Tokyo Japan.
- MMCBIR " Automatic Video Summarization" Itheri Yahiaoui, Bernard Merialdo et Benoit Huet, Inria/NFS Workshop on Multimedia Indexing (MMCBIR), September 2001, Paris, France.
- CBMI2001 " Optimal video summaries for simulated evaluation" Itheri Yahiaoui, Bernard Merialdo et Benoit Huet, European Workshop on Content-Based Multimedia Indexing (CBMI), September 19-21, 2001, Brescia, Italy.
- MMSP'01 " Automatic Summarization of Multi-episode Videos with the Simulated User Principle" Itheri Yahiaoui, Bernard Merialdo et Benoit Huet, Workshop on MultiMedia Signal Processing (MMSP'01), October 3-5, 2001, Cannes, France.
- EUSIPCO " Image Similarity for Automatic Video Summarization" Benoit Huet, Itheri Yahiaoui et Bernard Merialdo, 11th European Signal Processing Conference (EUSIPCO), September 3-6, 2002, Toulouse, France.
- IWDC2002 " Automatic Video Summarization" Bernard Merialdo, Benoit Huet, Itheri Yahiaoui et Fabrice Souvannavong, International Thyrrenian Workshop on Digital Communication, Advenced Methods for Multimedia

Signal Processing, September 8-11, 2002, Palazzo dei Congressi, Capri, Italy.

- ACM2002 "Generating TV Summaries for CE Devices" Gerhard Mekenkamp, Mauro Barbieri, Benoit Huet, Itheri Yahiaoui, Bernard Merialdo, Riccardo Leonardi and Michael Rose, ACM Multimedia 2002, December 3-5 2002, Juan Les Pins, France.

Conférences Nationales

- CORESA2000 " Résumés automatiques de séquences vidéo" Itheri Yahiaoui, Bernard Merialdo et Benoit Huet, Coresa, 19-20 Octobre 2000, Poitiers, France. CORESA2001 " Construction Automatique de résumés multi-vidéos" Itheri Yahiaoui, Bernard Merialdo et Benoit Huet, Coresa, 14-15 Novembre 2001, Dijon, France.
- RIAM "Construction et Evaluation automatique de résumés multi-vidéos" Itheri Yahiaoui, Bernard Merialdo et Benoit Huet, Analyse et Indexation Multimédia, June 20 2002, Université Bordeaux 1, France.
- GDR13&ISIS "User Evaluation of Multi-Episode Video Summaries" Itheri Yahiaoui, Bernard Merialdo et Benoit Huet, Indexation de documents et Recherche d'informations, GDR I3 et ISIS, July 9 2002, Grenoble, France.