République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Batna 2

Faculté des mathématiques et de l'informatique

Département d'informatique

Thèse

En vue d'obtention du diplôme de

Doctorat en Informatique

Environnement numérique de recherche d'information basé sur l'extraction d'information textuelle : Application aux vidéos de télé-enseignement

Présentée Par :

Belkacem Soundes

Devant le jury :

Président :	P[r] Bilami Azeddine	Professeur	Université de Batna 2
Rapporteur :	D[r] Guezouli Larbi	MCA	Université de Batna 2
Examinateurs :	P[r] Benmohammed Mohamed	Professeur	Université de Constantine 2
	D[r] Akhrouf Samir	MCA	Université de B. B. Arreridj
	D[r] Behloul Ali	MCA	Université de Batna 2
Invité :	D[r] Zidat Samir	MCA	Université de Batna 2

A mes chers enfants

Tesnime et Acil

Remerciement

Hamdo li Allah tout puissant de m'avoir accordé le pouvoir et la patience pour réaliser ce travail.

Je remercie vivement mon encadreur ***D^{r} Guezouli Larbi***, MCA à l'université de Batna 2 d'avoir accepté de m'encadrer, pour tous ces conseils, remarques et directives pour achever ce travail. Je le remercie aussi pour sa patience, son encouragement, sa compréhension et surtout pour sa confiance … merci infiniment.

Je remercie les membres de jury de m'avoir fait l'honneur d'accepté de juger mon travail. ***P^{r} Bilami Azeddine***, Professeur à l'université de Batna 2 en qualité de président de jury, ***P^{r} Benmohammed Mohamed***, Professeur à l'université de Constantine, ***D^{r} Akhrouf Samir***, MCA à l'université de B. B. Arreridj et ***D^{r} Behloul Ali***, MCA à l'université de Batna 2. Je remercie ***D^{r} Zidat Samir*** MCA à l'université de Batna 2 d'avoir accepté de m'encadrer en premier temps.

Je remercie mon marie, mes parents, mes frères et sœurs, mes oncles... sources de ma patience et puissance... Merci.

Merci à la famille ***SRI*** promo 2013 enseignants et étudiants…. Merci de me faire découvrir le monde de la recherche d'information et de traitement d'image, plein de merveilles.

Merci à tous ceux qui ont contribués à ce travail de près ou de loin. Merci aux ***P^{r} Haffad Tahar*** et ***P^{r} Lahbari Noureddine*** pour leurs conseils…merci.

Enfin, l'écriture d'un remerciement est toujours un moment d'émotion. Mais, revenir dans le temps et voir tous les gens qui étaient avec moi, qui mon aidé chacun de sa façon, est un grand plaisir ***Hamdo li Allah***.

Soundes

Résumé

A l'heure actuelle, les vidéos de présentations sont de plus en plus utilisées pour le télé-enseignement. Produites à termes quotidien, le nombre de vidéos disponibles en ligne ou archivées augmente explosivement. Le contenu de ces massives bases de données doit être facile à accéder et à rechercher dedans, ainsi, des systèmes automatiques d'indexation à base de contenu doivent être développés pour manipuler ce genre spécial de vidéos. Les performances de ces systèmes d'indexation à base de contenu sont fortement liées aux caractéristiques choisies pour décrire son contenu.

Dans cette thèse, nous nous concentrons autour du choix, de la détection et de l'extraction des caractéristiques visuelles permettant à la fois une bonne description du contenu ainsi que l'expressivité vis-à-vis l'utilisateur humain. Nous présenterons de nouvelles contributions relatives à la structuration et la description visuelle du contenu des vidéos de présentations.

Notre première contribution, se focalise sur l'organisation du contenu visuel des vidéos. Une méthode de segmentation récursive est mise au point pour identifier les différents segments des vidéos de présentation. La tâche est difficile à achever vue la nature longue et non structurée de ces vidéos, en plus de la nature des scènes homogènes et les différentes dégradations que présentent ce genre de vidéos. Pour une manipulation efficace, les trames sont échantillonnées et regroupées pour former des unités de traitement. Chaque segment subit deux types de traitement : l'extraction des caractéristiques, mesures de similarités et la vérification. Si une transition vers un autre segment est détectée, une procédure récursive de recherche de points de transitions est appliquée.

Notre deuxième contribution, s'articule sur l'extraction des informations textuelles à partir des vidéos de présentations. Le choix est justifié par les avantages majeurs que présentent ces informations par rapport aux autres éléments visuels. La région de projection de la diapositive est localisée en premier. Après, l'extraction et la segmentation du texte suivie d'une étape de filtrage est appliquée. L'utilisation des MPZ (Moments Pseudo-Zernike) montre une robustesse contre les conditions d'acquisitions.

Les méthodes proposées donnent de bons résultats vis-à-vis l'efficacité en termes de précision, avec un taux d'erreur minime, ainsi qu'en termes de temps de traitement. Ces méthodes ont présenté des résultats satisfaisants par rapport à des méthodes récentes.

Mots clés :

E-Learning, Recherche d'informations, Multimédia, Indexation par le contenu, Similarité, Segmentation vidéo, Extraction de texte, Moments Pseudo Zernike.

Abstract

At present, lecture videos are increasingly used for distance learning and the amount of lecture available online and archived videos is increase daily. The content of these massive databases must be easy to access and search in. Hence, automatic content-based indexing systems must be developed to handle this special kind of videos. Performances of every content-based indexing system are strongly related to relevance of its contents description. Therefore, the effective feature selection and extraction is a fundamental task.

In this thesis, we focus on the selection, detection and extraction of video visual characteristics allowing at the same time a good content description, also expressivity for human usage. We will present two new contributions related to the structuring and visual description of lecture videos contents.

Our first contribution surrounds on the organization of video's visual content. A recursive segmentation method is developed to identify the different segments of lecture videos. The task is hard to accomplish since this kind of videos is long and unstructured. For effective manipulations, frames are sampled and grouped together into processing units. Each segment will receive two types of treatment: feature extractions and verification. If a transition to another segment is detected, a recursive transition point detection procedure is applied.

Our second contribution consists on the extraction of textual information embedded within videos frames. In fact, scene text presents major advantages against other visual features. At first, slide projection region is located, and then text extraction and segmentation followed by a filtering step are applied. The use of PZM moments shows robustness against acquisitions conditions.

Proposed methods show good results in terms of accuracy, with a low error rate, also in terms of processing time. These methods have shown satisfactory results compared to recent methods.

Keywords

E-Learning, Information Retrieval, Multimedia, Content Indexing, Similarity, Video segmentation, Text extraction, Pseudo-Zernike Moments.

ملخص

في الوقت الحاضر، أشرطة الفيديو تُستخدم بشكل متزايد للتّعلم عن بعد. بشكل يومي يتم إنتاج عدد متزايد من فيديو المحاضرات سواء المتاحة عبر الإنترنت أو المحفوظة في أرشيفات. يجب أن يكون الوصول والبحث في محتوى هذه البيانات الضّخمة سهلا ودقيقا. وبالتالي، يجب تطوير أنظمة الفهرسة التّلقائية المستندة إلى المحتوى للتّعامل مع هذا النّوع الخاص من مقاطع الفيديو. يرتبط أداء كل نظام فهرسة قائم على المحتوى ارتباطًا وثيقًا بأهمية وصف محتوياته. ولذلك، فإنّ اختيار الميزة المناسبة والفعّالة واستخراجها مهمة أساسية.

في هذه الأطروحة نركز حول الاختيار والاكتشاف واستخراج خصائص بصرية تسمح في نفس الوقت بالوصف الجيد للمحتوى ولكن أيضًا سهلة للمستخدم البشري. سنقدم مساهمتين جديدتين تتعلقان بهيكلة ووصف المحتوى المرئي لمقاطع فيديو المحاضرات.

تتمثل أول مساهمة في تنظيم المحتوى البصري لمقاطع الفيديو. حيث تم تطوير طريقة تجزئة متكررة لتحديد أجزائه المختلفة. وهذا يعد أمرا صعبًا نظرًا لأن محتوى هذا النّوع من مقاطع الفيديو طويل وغير منتظم. لمعالجة فعالة، يتم أخذ عيّنات من إطارات الفيديو وتجميعها في وحدات؛ كل وحدة تخضع لنوعين من المعالجة: استخراج ميزة والتّحقّق. في حالة اكتشاف عملية انتقال إلى وحدة أخرى، فسيتم تطبيق إجراء الكشف المتكرر عن نقطة الانتقال.

مساهمتنا الثانية، تعتمد على استخراج المعلومات النّصية من مقاطع الفيديو. يتم تبرير استخدامهم من خلال المزايا الرئيسية التي تقدم ضد الميزات المرئية الأخرى. في البداية، يتم تحديد موقع منطقة إسقاط الشرائح، ثم يتم تحديد واستخراج النّصوص متبوعًا بخطوة تصفية. استخدام PZM يظهر متانة ضد ظروف تسجيل مقاطع الفيديو.

تُظهر المساهمات المقدمة نتائج جيدة من حيث الدّقة، مع معدل أخطاء منخفض، وأيضًا من حيث وقت المعالجة. أظهرت هذه الطّرق نتائج مرضية مقارنة بطرق اخرى.

كلمات مفتاحية

التّعلم الإلكتروني، استرجاع المعلومات، الوسائط المتعددة، فهرسة المحتوى، التّشابه، تجزئة الفيديو، استخراج النص، Pseudo-Zernike Moments.

Sommaire

Liste des figures

Liste des tableaux

Liste des algorithmes

Liste des abréviations

HU	Moments de Hu
MPZ	Moments Pseudo Zernike
MZ	Moments de Zernike
ROC	Reconnaissance Optique des Caractères
Stroke	Épaisseur de trait du caractère
SWT	Stroke Width Transform
Système EIT	Système d'extraction d'informations textuelles

Introduction Générale

Motivations et domaine d'application

A l'heure actuelle, l'adoption des documents multimédias pour l'apprentissage distant reçoit un intérêt croissant. Cette adoption est motivée essentiellement par la richesse d'informations issues des différents médias des vidéos éducatives, en plus de leurs capacités de présentation du contenu hétérogène sous forme simple. Récemment, ce genre de vidéos est de plus en plus produit et utilisé par l'enseignant ainsi que l'apprenant. Divers établissements pédagogiques, universités et centres de recherches exploitent ces documents pour le partage distant des connaissances. Ils adoptent les documents multimédias pour améliorer le processus d'apprentissage par l'enregistrement des cours et des différentes présentations académiques pour une utilisation ultérieure. Par la suite, les vidéos sont mises en ligne, afin que les étudiants puissent y accéder à tout moment et de n'importe quel endroit. Etant donné que les instructeurs tentent souvent d'utiliser des cours sous forme de présentations électroniques, ce genre de vidéos éducatives et communément reconnus par les vidéos de présentations. Des études récentes [1-3] ont montrées l'importance des vidéos à usage éducatif pour l'enseignement et le partage des connaissances (voir Fig.1).

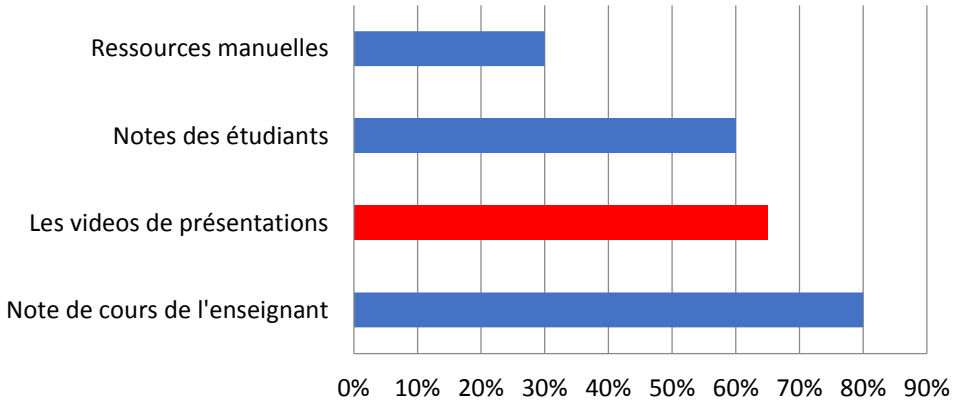

Fig.1. Importance des vidéos de présentation et d'autres ressources d'apprentissage

Les avantages, que présentent les vidéos de présentations, engendrent la construction de grandes archives des vidéos de présentations. Néanmoins, ce genre de vidéos reste loin d'être efficacement exploité et son utilisation reste toujours limitée. Ceci est dû, d'une part, aux problèmes et contraintes techniques que présentent les documents multimédias en général et d'autre part aux limitations que pose ce genre de vidéos comme le problème d'accès efficace au contenu et de recherche d'informations pertinentes. L'utilisateur se trouve souvent devant les questions suivantes :

- Comment trouver la vidéo pertinente dans une grande archive ?
- Comment trouver la position exacte de l'information à l'intérieur d'une vidéo ?

En effet, pour un utilisateur qui cherche une vidéo particulière, c'est impossible de passer par tous les documents multimédias existants surtout si leur nombre est important. Même une fois la vidéo est trouvée, ça reste toujours difficile de juger la pertinence de la vidéo seulement à partir des données descriptives insérées manuellement qui sont généralement abstraites et subjectives. Une autre difficulté réside dans le fait que lorsque la vidéo recherchée est trouvée, la position exacte de l'information désirée reste inconnue, ainsi, l'utilisateur doit voir et revoir la vidéo plusieurs fois pour trouver la position exacte. On se trouve ainsi devant une exigence des méthodes automatiques et efficaces permettant la recherche et l'accès aux vidéos de présentations à base de contenu.

Contributions

L'indexation à base de contenu est un domaine de recherche active [4-6]. Les systèmes existants commencent toujours par l'étape de segmentation pour l'organisation du contenu suivi de l'extraction des caractéristiques pour sa description.

Après une étude de la problématique et des méthodes existantes, nous nous sommes trouvés devant des questions pertinentes :

- Parmi les différents éléments visuels, lesquels répondent le mieux aux besoins de l'utilisateur ? Lequel décrit au mieux le contenu ?
- Quelles méthodes utilisées pour extraire l'élément sélectionné ?
- Comment segmenter et structurer le contenu des vidéos de présentations pour permettre l'indexation à base de contenu ?
- Les méthodes d'analyse du contenu des vidéos dites traditionnels peuvent-elles être adaptées pour l'utilisation avec les vidéos de présentations ?

La segmentation et structuration des vidéos pour l'organisation du contenu et l'extraction des trames clés est une étape essentielle de tout système d'indexation à base de contenu [7]. En effet, c'est la première étape de tout processus d'indexation. La segmentation vidéo consiste à l'identification des parties de la vidéo dites *segments* relativement à leurs genre et contenu. Contrairement aux vidéos ordinaires, les vidéos de présentations sont généralement non structurées et ne respectent pas la hiérarchie que présentent la plupart des vidéos (trame, segment, scène, ...). L'absence d'une structure ne permet qu'une recherche séquentielle. Par conséquent, la vidéo doit être revue plusieurs fois en avant et en arrière pour rechercher une information particulière, ce qui est inefficace et couteux. Pour prendre en charge l'indexation et la recherche automatique, la vidéo doit être segmentée en segments disjoints et significatifs. Un tel segment est appelé *segment diapositive* qui peut être considéré comme une séquence de trames contenant la même diapositive. Les segments sont délimités par des points de transitions marquants les frontières entre segments adjacents et marqués par le changement de la diapositive en cours.

Un point de transition entre segments diapositives est détecté lorsqu'une *transition de diapositive* s'est produite. Il s'agit d'une tâche très difficile à effectuer du fait que les vidéos de présentations sont généralement: 1) de faible qualité en raison des conditions d'acquisition tel que le bruit, la variation de la lumière, le mouvement de la caméra résultant de la

distorsion du perspective et de l'occlusion; 2) composition de scènes homogènes, où les changements dans la région de projection diapositive ne provoquent pas des changements significatifs de couleur; 3) mouvement de l'intervenant sur la région de projection diapositive, produisant un changement majeur pour la même diapositive; 4) les conditions d'acquisitions.

La détection des transitions diapositives a été étudiée dans de nombreuses recherches [5, 6], mais les solutions introduites fonctionnent dans des conditions prédéfinies et spécifiques, tel que la configuration stationnaire ou format multi-scènes, ce qui n'est pas le cas pour toutes les vidéos de conférences.

La détection des points de transitions (*Shot Boundary detection*) ou la segmentation temporelle des vidéos se base sur une dissemblance entre les trames adjacentes. Le processus de détection des points de transitions se compose généralement de trois étapes : l'extraction des caractéristiques visuelles, mesures de similarité entre les trames par paires et la détection. Dans la première étape, les caractéristiques visuelles sont extraites à partir des trames vidéo et des descripteurs sont générés pour les mesures de similarité. Parmi les caractéristiques visuelles utilisées, la forme est une caractéristique puissante adoptée par la plupart des systèmes d'indexation et recherche à base de contenu [3]. Les descripteurs de forme sont généralement divisés en deux catégories : ceux qui se basent sur le contour et d'autres qui se basent sur la région. Les approches basées sur le contour ne tiennent compte que de l'information sur les frontières de la région, tandis que les approches basées sur la région considèrent à la fois l'information sur le contour et l'intérieur de la région.

Les fonctions de moments sont les descripteurs de forme les plus utilisés. En fait, leur efficacité motive leurs larges utilisations dans les domaines de reconnaissance de formes et d'analyse d'images. Généralement, ils sont utilisés comme une caractéristique globale pour la description des régions. Il existe différents types de moments qui diffèrent principalement par leurs fonctions de base. Les études de Jeong et al. [7] montrent que les moments orthogonaux a base des polynômes orthogonaux tel que les Moments Pseudo Zernike MPZ [8] sont plus performants que les autres types des moments. Ils sont plus immunisés contre le bruit et plus expressifs avec un minimum de redondance d'information.

Les MPZ présentent d'autres avantages importants tels que la robustesse face aux bruits, à l'invariance de la rotation et aux changements d'échelle, ainsi que l'expressivité, l'extraction et la représentation des caractéristiques à plusieurs niveaux. Les MPZ ont été utilisés dans de

nombreuses applications telles que l'analyse d'images, la reconnaissance de formes, l'indexation et recherche à base de contenu, la reconnaissance faciale, le tatouage ...

La description du contenu pour l'indexation revient à extraire un ou plusieurs éléments qui peuvent décrire de manière robuste le contenu et servants pour la recherche et l'accès à la vidéo. L'élément sélectionné doit représenter au mieux le contenu, permettant d'identifier de manière claire la vidéo et aussi être facile à utiliser par l'utilisateur. Le contenu visuel de haut niveau peut être du texte, mouvement ou visage humain. Le texte est un élément important et peut servir pour décrire ou donner des informations contextuelles autour du contenu des vidéos. Selon les études de Zhu et al. et Yin et al. [9, 10] le texte est le premier objet capturé par l'œil humain et fournit les informations contextuelles les plus intuitives. En effet, le texte est un descripteur sémantique puissant : 1) il donne des informations contextuelles ; 2) il est expressif ; 3) il présente des infirmations non disponibles sur d'autres médias ; 4) il est plus facile à extraire, décrire et rechercher par rapport aux autres éléments visuels ; 5) il peut être utilisé dans plusieurs applications tel que l'indexation par le contenu ; et 6) il peut être considéré comme un objet et peut être utilisé dans la compression et le codage des documents vidéo. Les propriétés précédentes justifient la large utilisation du texte dans divers domaines : indexation à base de contenu des images et vidéos, l'analyse et l'indexation des vidéos de sport, détection des logos TV, l'identification des objets, détection des plaques de rue, détection des LANDMARKS et des véhicules etc. Dans le domaine éducatif, le texte est aussi utilisé par plusieurs applications, tel que le e-Learning, online Learning, et les visioconférences. Les performances des systèmes d'indexation sont directement liées aux taux d'extraction et de reconnaissance d'information textuelle.

Dans cette thèse, notre travail s'inscrit dans le contexte d'organisation et de description du contenu des vidéos de présentations pour l'indexation et la navigation à base de contenu. Notre axe de recherche s'étend sur les deux étapes les plus importantes de tous systèmes d'indexation : la segmentation et la description. Deux contributions ont été proposées, la première couvre la segmentation du contenu par la détection récursive des points de transitions, quant à la deuxième, elle permet l'extraction des informations textuelles enfouîtes dans les trames de la vidéo.

Dans la première contribution, nous proposons une approche pour l'analyse structurelle des vidéos de présentations. Le contenu des vidéos sera organisé sous forme de segments de manière automatique à l'aide d'une méthode récursive qui exploite la puissance des Moments

Pseudo Zernike pour regrouper les trames de contenu similaire et résultant d'un ensemble de trames clés.

La méthode de détection de points de transitions, se compose principalement de quatre étapes : échantillonnage et partitionnement des trames, détection et extraction de la région diapositive, extraction des caractéristiques et mesures de similarité, et enfin, détection récursive des points de transitions. L'échantillonnage et le partitionnement des trames consistent à considérer les trames dans un écart de temps prédéfini, cela évite les traitements inutiles. En effet, les trames consécutives ont des caractéristiques visuelles très similaires qui peuvent conduire à de fausses détections. Par la suite, l'ensemble des trames est partitionné en segments de *k* trames. Les traitements qui suivent considèrent les segments au lieu des trames échantillonnées, ceci optimise les performances et diminue la complexité du calcul. La similitude entre les segments successifs est calculée et comparée à un seuil donné. Si l'on trouve une dissemblance, le segment actuel est subdivisé récursivement et est vérifié pour localiser le point de transition exact. L'extraction des caractéristiques est critique pour la description et les mesures de similarité, ce qui justifie l'utilisation des MPZ invariants et expressifs.

Dans notre deuxième contribution [11], nous proposons une méthode de détection et de localisation des informations textuelles qui combine les propriétés des régions avec les caractéristiques du stroke pour détecter et segmenter les régions texte à partir des trames des vidéos de présentations. Deux caractéristiques efficaces sont combinées pour la détection et l'extraction de texte : (1) Les Moments Pseudo Zernike (MPZ), comme descripteurs de forme, avec (2) la propriété stroke extraite en utilisant l'opérateur local SWT (Stroke Width Transform) [12]. L'algorithme proposé se compose de trois étapes : segmentation de la région diapositive, détection et extraction des régions texte candidates et filtrage. La segmentation à base des MPZ est utilisée pour extraire la région diapositive ainsi que pour extraire le texte. La segmentation s'effectue par extraction de caractéristiques à l'aide des MPZ comme caractéristiques locales sur une fenêtre glissante, suivie d'une étape de classification en utilisant l'algorithme K-moyens. L'extraction de la région diapositive réduit la complexité du calcul et améliore les résultats du système en limitant les régions de recherche en une sous-image. Ça permet d'éliminer ainsi une grande partie de la région non textuelle. Ensuite, l'opérateur local SWT est utilisé pour regrouper les pixels candidats en composants connexes en fonction de la largeur du stroke (épaisseur de trait) auquel il appartient. Enfin, les composants sont filtrés à l'aide des propriétés géométriques et d'opérations morphologiques.

Les principales contributions de cette thèse sont les suivantes :

1 / L'identification robuste et efficace de *région diapositive* contenant la projection de la diapositive à partir des vidéos de basse résolution et avec une configuration non stationnaire. La récupération (détection et extraction) de la *région diapositive* est une étape importante du prétraitement, qui améliore les performances globales du système en réduisant le temps de calcul. La méthode proposée utilise des caractéristiques immunisées contre la variation de lumière et la distorsion de perspective avec un temps de calcul réduit.

2 / Extraction des caractéristiques robustes pour les mesures de similarité. Les MPZ sont invariants à l'échelle, à la rotation, à la distorsion de perspective et à la translation, ce qui rend la méthode robuste vis à vis les mouvements de la caméra, la distorsion de perspective et aux conditions de luminance.

3 / Méthode récursive de détection de points de transitions diapositives par subdivision récursive des segments en sous-segments plus petits. Par conséquent, la recherche du point de transition considérera les demi-segments à chaque fois et les comparaisons de similarité inutiles sont évitées. Deux seuils sont combinés : le seuil global est utilisé pour la vérification entre les segments et le seuil adapté local pour la recherche dans le segment.

4/ L'approche permet à la fois la détection et la segmentation d'informations textuelles sur des trames de faible résolution de manière directe et sans aucune étape de prétraitement. Cela aussi permettra son utilisation pour tout type de vidéos.

5/ Les résultats peuvent également être utilisés directement pour la reconnaissance optique des caractères.

6/ Le calcul rapide qui est très important pour les systèmes de recherche en temps réel, et même pour les systèmes d'indexations des grandes archives des données.

Organisation de la thèse

La thèse est organisée en deux parties : la première donne une introduction théorique de l'indexation par le contenu d'une manière générale, et plus spécifiquement, l'indexation des vidéos de présentation ainsi qu'un état de l'art concernant les différents modules d'un système d'indexation à base de contenu. Dans la deuxième partie nous introduirons nos contributions en détail.

La première partie est constituée de trois chapitres : Le premier introduit les systèmes d'indexation à base de contenu ainsi qu'un état de l'art des systèmes d'indexations des vidéos de présentations existants. Le deuxième chapitre parle de la segmentation des vidéos pour l'indexation où les différentes notions utilisées sont introduites, ainsi qu'un état de l'art des différentes méthodes existantes avec discussions des avantages et inconvénients. Le troisième chapitre concerne l'utilisation des informations textuelles pour la description visuelle ainsi que les différentes méthodes existantes avec discussion des points forts et faibles.

La deuxième partie est organisée en deux chapitres : le quatrième chapitre présente les détails de notre méthode récursive de segmentation temporelle des vidéos de présentations avec les résultats obtenus. Le cinquième chapitre présente notre méthode d'extraction des informations textuelles à partir des trames vidéo.

Une conclusion générale est présentée à la fin de la thèse présentant les résultats obtenus et donnant de futures perspectives.

Partie I

Chapitre 01

Les vidéos pour le télé-enseignement

Le télé-enseignement, permet la diffusion et le partage distant des connaissances. La production quotidienne du contenu pédagogique en résulte à une quantité massive des données. Néanmoins, ces ressources restent loin d'être efficacement exploitées.

L'exploitation efficace des informations vidéo disponibles est due à un accès efficace à leurs contenus. Les performances des systèmes d'indexation des vidéos à base de contenu, sont directement liées à l'efficacité et à la pertinence de la description des différents éléments visuels. Il est important d'extraire et de présenter une description des caractéristiques expressives et adéquates tout en respectant les particularités de vidéos éducatives.

Dans ce chapitre, nous introduirons le principe du télé-enseignement et les avantages et caractéristiques des vidéos de présentations. Les notions relatives à l'indexation à base de contenu seront présentées ainsi que les différentes caractéristiques visuelles existantes pour la description du contenu, en particulier les moments pseudo Zernike.

1. Les vidéos pour le télé-enseignement

1.1 Télé-enseignement

1.1.1. Notions Théoriques

Définition : Ensemble de connaissances fournies via différents médias : texte et images pour supporter le processus d'apprentissage distant. Ces connaissances délivrées sur un dispositif numérique : ordinateur ou appareil mobile, sont stockées et transmises via une connexion internet ou intranet [13].

D'une manière plus simple, le télé-enseignement est le processus de partager des connaissances nécessaires et utiles pour un apprenant distant via un dispositif électronique et une connexion pour un apprenant distant.

Les systèmes de télé-enseignement offrent un avantage principal qui est : facilité l'accès, le stockage et l'utilisation des différents médias pour l'apprentissage [14]. Ils permettent de rechercher et de trouver du contenu pédagogique sans obligation de déploiement. Deux modes d'utilisation peuvent être utilisés : *synchrone* si les apprenants doivent être connectés en même temps et interagissent en temps réel ou *asynchrone* si les ressources pédagogiques sont disponibles à l'apprenant distant sans conditions sur le temps d'accès ni sur le temps d'interaction [15].

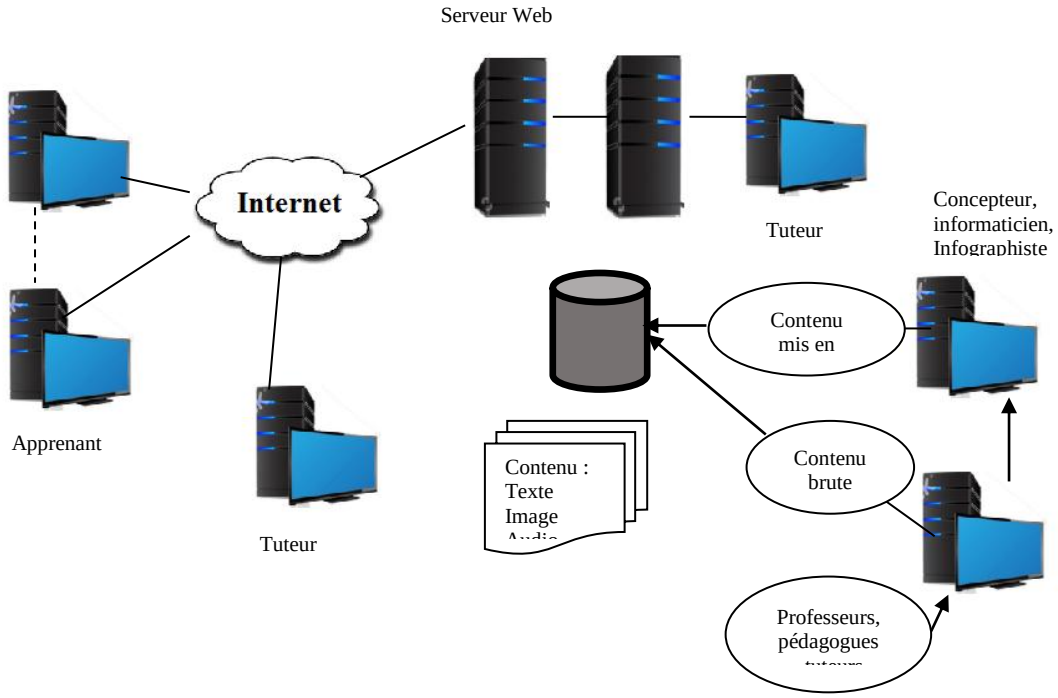

Fig. 1.1. Architecture générale des systèmes de télé-enseignement [4].

1.1.2. Les vidéos pour le télé-enseignement

Les vidéos à contenu pédagogique ou encore les vidéos de présentations sont un outil important pour l'apprentissage et la formation à distance [13-18]. L'organisation des connaissances sous cette forme permet d'approfondir le processus l'apprentissage de l'apprennent en lui présentant l'information d'une manière plus organisée et permettre de faire des liaisons entre différentes source d'informations. En plus, pour le cas du télé-enseignement, cela offre la possibilité d'y revenir à tout moment et, par conséquent, revoir quand c'est nécessaire.

Les documents multimédias comportent plusieurs médias : texte, image et audio. Chaque média est considéré comme une source d'informations. Le point fort des documents multimédias résident dans leurs richesses d'informations par rapport aux autres documents. En effet, l'utilisation des informations hétérogènes issues des différents médias visuel et audio sont à l'origine de sa richesse et expressivité sémantique. Ce qui justifie leurs larges utilisations dans le domaine de l'enseignement en général.

Les documents multimédias, permettent d'avoir plusieurs sources d'informations issues des différents médias. Les informations qui peuvent être extraites à partir d'une vidéo sont [19, 20] : les métadonnées ajoutées manuellement, les informations audio à partir d'un média audio et les informations visuelles issues des trames vidéos.

- *Les métadonnées :* ajoutées manuellement lors de la création de la vidéo telle que le titre, la date de création, le format, la taille et éventuellement une description. Néanmoins, ces informations restent abstraites, subjectives et consomment beaucoup de temps et ressources.
- *Audio et Paroles :* issue du signal audio via des méthodes de reconnaissance de voix.
- *Contenue visuel :* le contenu visuel des trames.
- *Autres sources d'informations* : sont les manuscrits utilisés lors de la présentation et les supports de cours aux format électronique dans le cas des vidéos de présentations [21].

La richesse et diversité de contenu sont les deux avantages qui motivent l'utilisation des vidéos de présentation pour la diffusion et le partage distant des connaissances. En plus, les technologies d'acquisition et de transmission ont facilité leurs acquisitions et mise en ligne, et par conséquent, la construction de grandes archives qui doivent être efficacement stockées, indexées et recherchées par leurs contenus. Cependant, l'absence de systèmes de recherche

efficaces et robustes pour ce genre de vidéos est une limitation majeure qui reçoit à l'heure actuelle un intérêt croissant.

1.1.3. Caractéristiques des vidéos de présentation

Les vidéos de présentation présentent plusieurs difficultés pour l'utilisateur ainsi que pour les systèmes d'indexation à base de contenu [22]. Ce genre de vidéos se caractérise par la fusion des propriétés des documents multimédias en général dites *ordinaires,* en plus des contraintes imposées par leur nature spécifique fortement liée aux conditions d'acquisition et à l'organisation de leurs contenus [2, 5, 23] :

1. *La composition homogène des scènes :* Le changement de trame ne provoque pas un changement significatif dans la région de projection ;
2. *La basse qualité due aux conditions d'acquisition* (influence de l'environnement extérieur et enregistrement non professionnel) ;
3. *Le contenu non structuré :* et par conséquent rend la segmentation une tâche délicate. Il est important de noter que même pour la segmentation manuelle ça reste difficile et subjective ;
4. *La diversité de contenu :* le contenu d'une même vidéo ou d'un même corpus est de nature hétérogène. Ainsi, différents styles de présentation sont utilisés, différents intervenants et différents types de présentations coexistent et doivent être manipulés par les mêmes outils ;
5. *La présence de ressources supplémentaires :* un nombre important des méthodes proposées pour l'indexation de ce genre de vidéos exploitent des ressources supplémentaires pour la structuration de contenu par une synchronisation entre la présentation et le support de cours sous format électronique. Néanmoins, la présence de telles ressources supplémentaires n'est pas toujours possible, et dans le cas général, les vidéos sont fournies seules.

En conclusion, les systèmes d'indexation et de recherche de vidéos de présentations à base de contenu doivent surélever les défis suivants :

1. Contenus hétérogènes issus des différents médias et de différentes sources.
2. Le contenu non structuré.
3. La mauvaise qualité causée par les conditions d'acquisitions.

1.1.4. Les types des vidéos de présentations

Les vidéos de présentations peuvent être catégorisées selon : l'intervenant, la configuration de la caméra, l'enregistrement et la diffusion.

L'intervenant peut utiliser différents types de régions d'intérêt : des diapositives, le tableau, ou une vue web du cours. Il peut être présent ou non dans les trames ainsi que la possibilité de faire apparaitre ou non l'audience.

La configuration de la caméra utilisée est une caractéristique importante qui doit être prise en considération par toute méthode de segmentation ou description du contenu. En effet, le mouvement de la caméra est une condition défavorable, ce mouvement engendre des vidéos de basse qualité, distorsion de perspectives et changements de vue et par conséquent influe négativement sur les performances de ces méthodes. Deux types de configurations sont considérés : configuration stationnaire et configuration non stationnaire. Si la position de la caméra est fixe le long de la vidéo on parle d'une *configuration stationnaire* ; dans le cas contraire on parle d'une *configuration non stationnaire.*

(a) *(b)*

Fig. 1.2. Exemple de types des vidéos de présentations. *(a) Vidéos à scènes unique, (b) Vidéos Multi-scènes* [21].

Selon le mode de diffusion et d'enregistrement, deux catégories de vidéos sont obtenues : *Vidéos à scènes uniques* et *Vidéos Multi-scènes*.

a) Vidéos à scènes uniques : Une seule caméra est utilisée pour générer un flux vidéo unique. Ce genre de vidéos peut adopter une configuration stationnaire ou non. Généralement fournit sans métadonnées supplémentaires. Les trames de ce genre incluent l'intervenant et/ou l'audience et/ou le tableau ou région diapositive. Il s'agit de la forme la plus populaire mais aussi celle qui présente plus de contraintes.

b) Vidéos Multi-scènes : Combinent deux flux vidéo : un flux pour la région diapositive et un autre pour l'intervenant [5]. Dans ce cas, deux caméras stationnaires sont utilisées. Ce

type de vidéos considère les avantages de configuration stationnaire avec la région de projection bien définie et capturée. Mais, généralement, ce n'est pas le cas des vidéos de présentations. (Fig. 1.2**)**

Dans ce travail, nous visons à traiter les vidéos de présentation à scène unique, sans données supplémentaires avec une configuration non stationnaire. Nous optons pour ces conditions difficiles afin de proposer des solutions globales pour les différentes manipulations à base de contenu des vidéos de présentation.

1.2 Structure générale des systèmes d'indexation des vidéos à base de contenu

1.2.1. L'analyse Structurelle

Peut être définie [4] par la segmentation d'une vidéo a un nombre d'éléments structurels qui ont une sémantique commune. Ce processus de segmentation comprend la détection des points de transitions, l'extraction des trames clés et la segmentation en scènes.

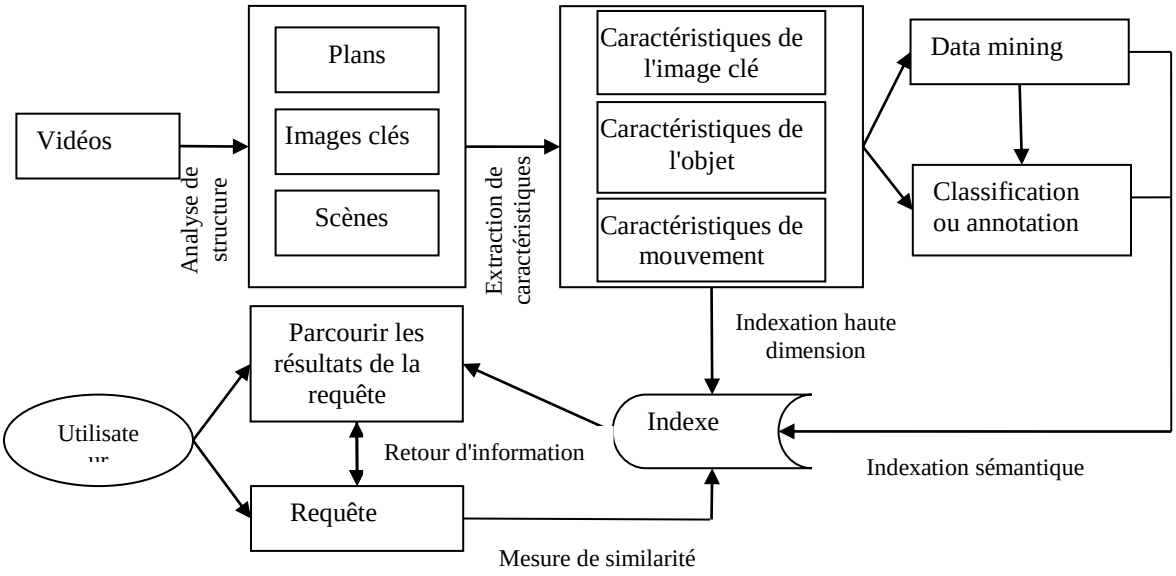

Fig. 1.3. Cadre générique pour l'indexation et recherche des vidéos basée sur le contenu visuel [4].

1.2.1.1. La détection des points de transitions

C'est le processus qui permet de détecter une distance entre les trames adjacentes (dissimilaires) en utilisant une mesure de similarité [24]. Ceci permet de segmenter la vidéo

en un ensemble de segments ayant des trames a contenus similaires séparés par des points de transitions.

1.2.1.2. L'extraction des trames clés

Les trames d'un même segment présentent une grande redondance d'informations [25]. Les trames sont fortement liées à un sujet particulier et partagent un grand nombre de caractéristiques visuelles. Par conséquent, tout segment peut être représenté par une trame particulière (trame clé) qui présente au mieux les caractéristiques communes.

1.2.2. L'extraction des caractéristiques pour l'indexation à base de contenu

La segmentation des vidéos permet l'extraction des unités de traitement des segments (*shots*). Les caractéristiques extraites à partir de ces segments sont utilisées pour la description, l'annotation et l'indexation. Ainsi, la sélection des caractéristiques appropriées et les choix des méthodes robustes pour l'extraction est une tâche critique dont dépond les performances du système d'indexation à base de contenu.

L'analyse du contenu d'une vidéo peut se faire à deux niveaux : le bas niveau et le haut niveau (sémantique).

1.2.3. L'indexation

L'étape d'indexation consiste à segmenter les documents vidéo en segments. A partir de ces segments on extrait un sous ensemble de trames clés. Leurs contenus sont caractérisés et codés sous format approprié.

L'extraction et la description effective du contenu est une étape importante de tout système d'indexation [26]. Les caractéristiques extraites sont groupées suivant un modèle mathématique pour la construction d'index [19]. L'extraction et le codage des caractéristiques vidéos doit prendre en considération la sémantique entre le contenu des documents multimédias et l'utilisateur [20]. Contrairement aux documents textuels comprenant des informations explicites, les informations sémantiques des vidéos sont enfouies dans leurs contenus et réparties sur les différents médias. Par conséquent, une bonne indexation revient à rapprocher le contenu aux expressions et perception humaines.

De point de vue utilisateur, il est plus pratique, facile et commode de faire revenir l'indexation multimédia à une indexation textuelle. L'extraction des caractéristiques permet

d'extraire un ensemble de termes significatifs représentant le contenu des documents. Ces termes constituent un index auquel seront comparées les requêtes des utilisateurs. Néanmoins, l'extraction de ces caractéristiques est une tâche délicate à cause des caractéristiques des vidéos.

1.2.4. L'interrogation

L'interrogation consiste en trois étapes [4] : La formulation des besoins sous forme d'une requête, la recherche des informations dans l'index en utilisant une mesure de similarité et la présentation des résultats obtenus.

1.2.4.1. La formulation

L'utilisateur exprime ses besoins sous forme d'une requête en respectant un formalisme imposé par le système. Il existe plusieurs types de requêtes : requête par mot clés, par exemple, par prototype ou par objet. On peut utiliser une combinaison d'une ou plusieurs types de requêtes. L'utilisation d'une requête à mots clés consiste à utilisées des termes significatifs représentants les contenus désirés.

1.2.4.2. Mesures de correspondance

La pertinence des documents par rapport à la requête est mesurée en utilisant une mesure de correspondance.

1.2.4.3. Présentation des résultats

Les documents ayant les plus grands scores de correspondances sont fournis à l'utilisateur souvent sous forme d'une liste ordonnée d'une manière décroissante selon leurs pertinences.

1.2.5. Indexation à base de contenu des vidéos de présentations

Les systèmes d'indexation à base de contenu des vidéos de présentations peuvent être regroupés selon les caractéristiques utilisées pour l'indexation en trois catégories : ceux à base des métadonnées, ceux à bases des informations textuelles et ceux à bases des informations audio.

Les systèmes à base des métadonnées insérées manuellement telle que Berkeley webcast et Open Yale course produisent des résultats inefficaces et leur utilisation est limitée.

D'autres méthodes à base d'informations audios ont été proposées, tel que MIT OpenCourseWare, Speech@FIT, NTU qui permettent de faire l'indexation des vidéos à travers des informations extraites du flux audio via des méthodes de reconnaissance vocale. La recherche se fait via des requêtes textuelles. Néanmoins, comme cité précédemment, l'audio dans les vidéos de présentations est généralement non structuré et spontané. En plus, il ne permet pas de respecter l'ordre chronologique des diapositives et par conséquent une segmentation temporelle ne peut avoir lieu.

Les systèmes exploitants les informations visuelles, en particuliers le texte via des outils de reconnaissance optique des caractères, présentent des résultats meilleurs que celles produites par les autres catégories. Wang et al. [27] propose un système d'indexation et de segmentation en utilisant la reconnaissance optique des caractères vidéo (ROCV). La segmentation se fait via une synchronisation entre le contenu des diapositives et les présentations en format électronique. Le texte est détecté par le calcul du ratio entre le texte et l'arrière-plan. Talkminer [28] est un système de recherche dans les vidéos de présentations. Il exploite les données insérées manuellement avec ceux résultants des ROCV pour permettre la recherche. Néanmoins, l'absence d'une étape de détection et de localisation de texte donne des performances assez satisfaisantes. Sack et al.[29] exploitent les notions sémantiques pour l'indexation. Différentes entités sont extraites auxquelles sont attribués des scores. Néanmoins, la définition de l'entité en elle-même et quel score lui attribué est un problème difficile à résoudre. Yang [5] propose des méthodes de segmentation et d'extraction d'information textuelle à partir des vidéos multi-scènes via des propriétés de contours. Gigonzac et al. [30] utilisent le modèle de Markov pour extraire et décrire la séquence des diapositives. Adcokc et al. [19][28] utilisent le classifieur SVM pour regrouper les différentes trames en trois catégories : diapositive/intervenant/arrière-plan, puis appliquer les techniques de ROCV sur une partie au lieu de l'appliquer à la vidéo entière. Tuna et al. [18] extraient les trames clés à partir desquels on extrait le texte. Plusieurs méthodes d'amélioration et de transformation d'images ont étaient exploitées pour améliorer le taux de reconnaissance, mais n'opèrent pas pour des contenus complexes. Shah et al. [38] utilisent SVM pour la segmentation, le suivie et l'extraction des informations à base d'approches linguistique. Furini et al. [2] exploitent les informations issues du suivi social pour estimer la l'importance des vidéos entre intéressant, significatif et autre afin d'améliorer les résultats d'indexation.

Des approches dites hybride, comme les travaux de Balasubramanian et al. [23] et Yang et al. [5], combinent des caractéristiques multi-modèles pour l'indexation issue du média audio et des informations textuelles.

1.3 Description visuelle d'une image

L'analyse du contenu d'une vidéo peut se faire à deux niveaux : le bas niveau et le haut niveau (sémantique).

1.3.1. Caractéristiques de bas niveau

Les caractéristiques de ce niveau concernent le niveau signal du document et sa représentation numérique : couleur, texture et forme. La plupart des systèmes d'indexation utilisent des caractéristiques de bas niveau mais ils présentent plusieurs limitations :

a) Loin de l'utilisateur ce qui influe sur l'efficacité et le rendement du système.

b) La richesse des documents multimédias est non exploitée du fait qu'un seul signal est utilisé à la fois.

c) Ne marche que pour les recherches à base d'exemple, alors que ce dernier n'est pas toujours présent pour l'utilisateur.

1.3.1.1. La texture

Les descripteurs de texture tels que DCT (Discret Cosine Transform) et LBP (Local Binary Pattern) sont exploités pour l'extraction des informations textuelles. Mais, ce genre de méthodes présente de faibles résultats avec les vidéos de présentations. En effet, ces méthodes ne peuvent pas séparer le texte des autres objets texturés de la scène et par conséquent non adéquat à l'utilisation avec ce genre de vidéos [2].

1.3.1.2. La forme

La forme permet d'identifier des objets tel qu'ils sont dans le monde réel. Les descripteurs de forme varis entre : contours et régions.

Les approches existantes exploitent les caractéristiques de contour pour décrire la forme. Ces approches n'exploitent que les informations du contour des formes [31], contrairement à celles basées sur les régions qui extraient les informations de toute la région. Mais le temps de calcul est un inconvénient majeur des méthodes de cette approche.

Les approches à base de régions, impliquent tous les pixels à l'intérieure de la région pour l'extraction des caractéristiques [32]. Ils exploitent plusieurs types de moments pour la description des régions [33].

1.3.2. Caractéristiques de haut niveau

A la recherche d'un document, l'utilisateur souvent exprime ces besoins de façon simple et proche a sa langue naturelle sous forme de mots ou paroles. Mais l'extraction de ces caractéristiques est plus difficile et plus couteuse que celles de bas niveau, mais l'utilisateur est plus alaise avec le système ainsi que les résultats sont plus pertinents.

1.3.2.1. Métadonnées manuelles

La plupart des systèmes d'indexation vidéo à base de contenu utilisent des métadonnées insérées manuellement aux vidéos. Mais ces informations sont :

- Généralement subjectives [23] ;
- Abstraites et ne donnent que des informations très générales et de haut niveau ;
- Insérées manuellement par l'être humain, ce qui est couteux [5].

1.3.2.2. L'audio

L'audio en général varie entre parole, musique, etc. Les systèmes d'indexation utilisent la parole comme caractéristique. Ils exploitent des outils de reconnaissance vocale pour générer des informations textuelles à partir du flux audio.

A présent, les systèmes existants basés sur l'audio donnent des résultats assez satisfaisants mais restent inférieurs résultats obtenus en utilisant d'autres caractéristiques. Ceci est dû, d'une part au bruit et d'autre part à la qualité des vidéos utilisées. Il reste à noter qu'il s'agit toujours d'une indexation textuelle.

Dans les vidéos éducatives, qui sont généralement produites par des non-professionnels, manquent de structure et de pré-préparation. Le discours produit par l'intervenant est généralement spontané, non structuré et comprend beaucoup de mots hors lexique et de longs silences. En plus, ce genre de vidéos qui sont enregistrées dans des scènes naturelles comprend beaucoup de bruit qui influe négativement sur le taux de reconnaissance de la parole [2].

1.3.2.3. Le Texte

Les informations visuelles peuvent être exploitées pour l'indexation. Ces éléments varies entre : le texte, le mouvement et le visage humain [10]. Parmi ces éléments, le texte présente des caractéristiques importantes et motivent sa large utilisation pour l'indexation à base de contenu :

- Principale source d'émission et de transmission d'informations ;
- Le texte lu ou écrit est rapide à extraire et effectif pour la description de haut niveau ;
- Le texte des diapositives est utilisé comme sommaire du discours.

Les différentes techniques, caractéristiques et avantages relatifs à l'extraction et l'indexation à base de texte seront introduites en détails dans le troisième chapitre.

1.4 Les descripteurs de formes

1.4.1. Définitions

Généralement, les moments calculés pour une image de dimension N*M sont calculés via l'équation suivante :

$$\boldsymbol{M = \varphi(x, y) \cdot f(x, y)} \qquad (1.1)$$

où :

- $\varphi(x, y)$ Fonction de base (noyau).
- $f(x, y)$ Fonction d'intensité *f* au pixel *(x,y).*

1.4.2. Type des moments

Le type des moments varie suivant la fonction de base utilisée. On distingue deux types : *orthogonaux* qui utilisent des polynômes orthogonaux et autres *non orthogonaux* le cas où des polynômes non orthogonaux sont utilisés comme noyau.

1.4.2.1. Moments orthogonaux

Selon les polynômes de base utilisés, différents types de moments sont obtenus et généralement portent le même nom de ces polynômes. En 1980, Teague [34] a introduit les moments de Zernike, un ensemble des moments invariants à base des polynômes de Zernike.

Une autre version modifiée de ces moments est proposée en 1988 par Teh [8] à base des polynômes pseudo Zernike.

Des études [35] ont montrées l'efficacité de description des moments orthogonaux par rapport aux autres types des moments. Les moments orthogonaux sont utilisés comme caractéristiques sensibles pour la classification et les applications de reconnaissance [36]. En effet, la propriété d'orthogonalité réduit la sensibilité au bruit, la redondance d'information et offre la possibilité d'une description multi-niveaux. En plus, l'orthogonalité engendre des moments invariants à la rotation et à l'échelle.

Les moments pseudo Zernike produisent de meilleurs résultats par rapport aux autres types de moments orthogonaux [35]. Les expérimentations ont montré que les MPZ sont plus robustes au bruit et plus expressives que les moments de Zernike [37, 38].

1.4.2.2. Moments non orthogonaux

Les moments géométriques d'ordre p et répétions q pour une image de taille N*M, ayant la fonction d'intensité *f(x,y)* par la formule suivante :

$$M_{p,q} = \iint x^p x^q f(x,y)dxdy_D \; ; p; q = 1,2,,.. \tag{1.2}$$

Pour assurer l'invariance à la translation, les moments centraux sont calculés par rapport au centre de l'image (x_c, y_c) :

$$\mu_{p,q} = \sum_{x,y}(x - x_c)^p (y - y_c)^q f(x,y) \tag{1.3}$$

Tel que : (x_c, y_c) sont les coordonnées du centre de l'image.

$x_c = {M_{10}}/{M_{00}}$; $y_c = {M_{01}}/{M_{00}}$; $M_{ij}=\sum\sum x^i y^j f(x,y)$;

L'invariance à l'échelle est assurée par la normalisation des moments centrés :

$$\eta_{p,q} = \frac{\mu_{p,q}}{\mu_{0,0}^{\gamma}} \tag{1.4}$$

Tel que $\gamma = \frac{p+q}{2} + 1, \;\; \forall\, p + q \geq 2$

En 1962, HU [39] a introduit un ensemble de moments invariants pour caractériser l'image. Il calcule les moments d'ordre P_{max} = 3 sur une image *N* * *M*. Ainsi, sept moments sont invariants aux directions, échelles et orientations. En utilisant les équations suivantes :

$$I_1 = \eta_{20} + \eta_{02} \tag{1.5}$$

$$I_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}{}^2 \tag{1.6}$$

$$I_3 = (\eta_{30} - 3\eta_{12})^2 + (\eta_{21} - 3\eta_{03})^2 \tag{1.7}$$

$$I_4 = (\eta_{30} - \eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{1.8}$$

$$I_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} - \eta_{12})[(\eta_{30} - \eta_{12})^2 - 3(\eta_{21} - \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} - \eta_{03})[3(\eta_{30} - \eta_{12})^2 - (\eta_{21} - \eta_{03})^2] \tag{1.9}$$

$$I_6 = (\eta_{20} - \eta_{012})[(\eta_{30} - \eta_{12})^2 - (\eta_{21} - \eta_{03})^2 + 4\eta_{11}(\eta_{30} - \eta_{12})(\eta_{21} - \eta_{03})] \tag{1.10}$$

$$I_7 = (3\eta_{21} - \eta_{03})(\eta_{30} - \eta_{12})[(\eta_{30} - \eta_{12})^2 - 3(\eta_{21} - \eta_{03})^2] + (\eta_{30} - 3\eta_{012})(\eta_{21} - \eta_{03})[3(\eta_{30} - \eta_{12})^2 - (\eta_{21} - \eta_{03})^2] \tag{1.11}$$

Les six premiers moments décrivent la forme en assurant une invariance à la translation, à l'échelle et à la rotation. Le septième moment assure une invariance da la distorsion géométrique [40].

1.4.3. Les moments pseudo Zernike

1.4.3.1. Notions théoriques

Les moments pseudo Zernike MPZ sont des moments orthogonaux à base de polynômes Pseudo Zernike définis dans le système polaire. Le moment PZ d'ordre p et répétition q, calculé pour une image de taille N * N ayant la fonction d'intensité $f(x, y)$ est donné par l'équation suivante :

$$PZM_{p,q} = \frac{p+1}{\pi} \iint_{x^2+y^2 \leq 1} V_{p,q}^*(x, y) f(x, y) dx dy \tag{1.12}$$

où $V_{p,q}^*(x, y)$ est le conjugué complexe du polynôme pseudo Zernike $V_{p,q}(x, y)$.

$$V_{p,q}(x,y) = R_{p,q}(r)e^{jq\theta} \quad (1.13)$$

Tel que :

- $R_{p,q}(r)$: Polynôme radial sur les coordonnées polaires *(r, θ).*
- $e^{jq\theta}$: Fonction angulaire, tel que : $e^{jq\theta} = (\cos\theta + j\sin\theta)^q$.
- p: Ordre du moment, $p \geq 0$.
- q: Répétition du moment, $0 \leq |q| \leq p$. Seules les valeurs positives sont utilisées puisque les valeurs négatives peuvent être obtenues en utilisant le conjugué complexe : $PZM_{p,-q} = PZM^*_{p,q}$.
- j : Nombre imaginaire $j = \sqrt{-1}$.
- θ : Angle entre r et l'axe des X. $\theta = \tan^{-1}(x/y)$ et $\theta \in [0, 2\pi]$.
- r : Longueur du vecteur de l'origine $(\bar{x}, \bar{y})$ jusqu'au pixel (x, y). $r = \sqrt{x^2 + y^2}$.
- $R_{p,q}(r)$ est donné par :

$$R_{p,q}(r) = \sum_{s=0}^{p-|q|}(-1)^s \frac{(2p+1-s)!}{s!(p+|q|+1-s)!(p-|q|-s)!} r^{p-s} \quad (1.14)$$

La forme discrète des moments PZ est donnée par :

$$PZM_{p,q}(f(x,y)) = \frac{p+1}{\pi}\sum_{i=1}^{N}\sum_{j=1}^{N} V^*_{i,j}(x,y)f(x,y)\,\Delta x\Delta y \quad (1.15)$$

Tel que : $\Delta x = \frac{N}{2}$; $\Delta y = M/2$.

La reconstruction d'image est possible via l'équation :

$$f(x,y) = \sum_{i=0}^{p}\sum_{j=-p}^{q} PZM_{i,j}V_{i,j}(x,y) \quad (1.16)$$

Les moments PZ sont définis sur des coordonnées polaires dans un cercle unitaire ; alors que les pixels de l'image carrée doivent être normalisés à l'intervalle [0,1] où $x^2 + y^2 \leq 1$. [41]

La normalisation est effectuée par une transformation linéaire des coordonnées des pixels en un système polaire, où le centre de l'image est pris comme origine du cercle unité. La normalisation peut être effectuée de deux façons : [32]

a. ***Normalisation du cercle d'unité dans l'image :*** Le cercle d'unité est mappé dans l'image. Néanmoins, les pixels en dehors du cercle sont ignorés et ne seront pas pris en compte pour le calcul des moments PZ.

b. ***Normalisation de l'image dans le cercle d'unité :*** L'image est mappée a l'intérieure du cercle d'unité. Par conséquent, la totalité de l'image est incluse dans le calcul des moments sans aucune perte d'informations.

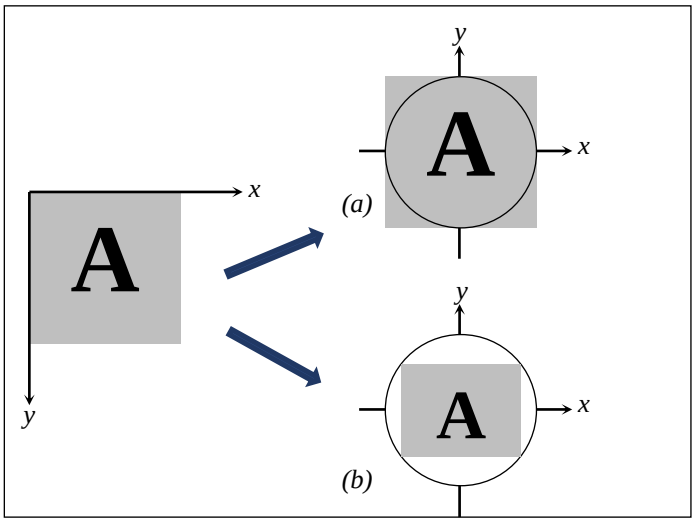

Fig. 1.4. Méthodes de normalisation d'image pour le calcul des moments PZ. ***(a)*** **Cercle dans l'image,** ***(b)*** **Image dans le cercle.**

Pour éviter la perte d'informations, la deuxième méthode de normalisation est généralement utilisée. Les pixels normalisés (x_c,y_c) sont obtenus par :

$$x_c = \frac{2x+1-N}{N\sqrt{2}} \tag{1.17}$$

$$y_c = \frac{2y+1-N}{N\sqrt{2}} \tag{1.18}$$

Tel que : (x,y) sont les coordonnés avant la normalisation.

1.4.3.2. Caractéristiques

Les moments PZ présentent plusieurs caractéristiques intéressantes :

1) Invariance à la rotation [42] ;
2) Moins sensibles au bruit qu'autres moments orthogonaux [41] ;

3) Expressivité : redondance d'information minime [32]. La propriété *orthogonalité* des polynômes Pseudo Zernike résulte en une redondance de l'information proche de zéro ;

4) Représentation multi-niveaux : Les moments de différents ordres se réfèrent à des caractéristiques différentes. Les moments de faibles ordres capturent les détails généraux, quant à ceux d'ordre supérieure, ils capturent plus d'informations locales [43] ;

5) Effectivité : Le nombre de moments calculés pour un ordre et plus que celui d'autres moments orthogonaux, tel que les moments de Zernike.

6) Reconstruction d'image [43].

Les MPZ sont les moments les plus utilisés pour la description dans les applications de reconnaissance de forme [37] et d'analyse d'image. Les PZM sont utilisés dans de nombreux domaines tels que la reconnaissance optique de caractères, la classification de motifs, la reconnaissance faciale, la récupération d'images basée sur le contenu, le tatouage d'image, la reconstruction d'image, la vérification et la reconnaissance de signatures.

1.4.3.3. L'ordre des moments

Le nombre des moments extraits est relatif à l'ordre de moments utilisés. Différents ordres décrivent différentes caractéristiques où les moments d'ordre inférieur décrivent les caractéristiques globales de l'image, et ceux d'ordres supérieurs donnent plus d'informations locales. La figure Fig. 1.5 donne un exemple d'influence du choix d'ordre du calcul des moments PZ sur la qualité de description et par conséquent celle de reconstruction. Comme montré sur la figure, l'ordre p_{max}=12 permet d'obtenir une bonne qualité de description des caractères à partir d'une image à haute résolution.

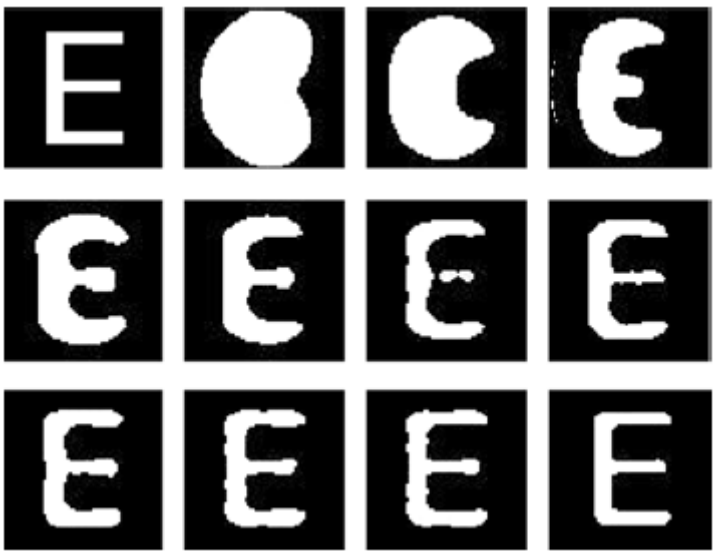

Fig. 1.5. Image reconstruite du caractère E à partir d'ordre p_{max}=2 jusqu'à p_{max}=12 [43].

Le nombre de moments calculés pour l'ordre p_{max} est égale à $(1+ p_{max})^2$. La construction du vecteur caractéristique de l'image est donnée par :

$$PZM_{Pmax}(f(x,y)) = \{|PZM_{p,q}(f(x,y))|, p = 1..p_{max}, q = 1..p\} \quad (1.19)$$

Les vecteurs sont construits à base de la magnitude des MPZ. D'une part, elle est invariante à la rotation, et d'autre part elle permet de considérer la moitié des moments calculés pour la description des caractéristiques du fait que :

$$PZM_{p,-q}(f(x,y)) = PZM_{p,-q}{}^{*}(f(x,y)) \quad (1.20)$$

$$|PZM_{p,-q}(f(x,y))| = |PZM_{p,q}{}^{*}(f(x,y))| = |PZM_{p,q}(f(x,y))| \quad (1.21)$$

où seulement la moitié du nombre des moments est pris en considération pour la construction du vecteur caractéristique. Le tableau ci-dessous montre un exemple des moments prisent comme caractéristiques pour $p_{max}=5$.

N	**PZM**	**Nombre**
0	$PZM_{p,q}$	1
1	$PZM_{p,q}; PZM_{p,q}$	2
2	$PZM_{p,q}; PZM_{p,q}; PZM_{p,q}$	3
3	$PZM_{p,q}PZM_{p,q}; PZM_{p,q}; PZM_{p,q}$	4
4	$PZM_{p,q}; PZM_{p,q}; PZM_{p,q}; PZM_{p,q}; PZM_{p,q}$	5
5	$PZM_{p,q}; PZM_{p,q}; PZM_{p,q}; PZM_{p,q}; PZM_{p,q}; PZM_{p,q}$	6

1.5 Conclusion

Les vidéos de présentations sont une ressource riche d'informations importées à partir des différents médias. L'exploitation de ce genre de vidéos pour le télé-enseignement permettra d'améliorer le processus d'apprentissage pour l'apprennent.

L'indexation à base de contenu permet l'accès et la recherche efficaces dans les grandes archives multimédias. Ces systèmes se basent sur la bonne description et organisation du contenu. Les méthodes les plus efficaces se basent sur les propriétés internes des régions des trames. Parmi les descripteurs existants, les moments pseudo Zernike présentent plusieurs points forts et par conséquent motive leur utilisation pour la description du contenu des vidéos de présentations. Néanmoins, une bonne description nécessite une organisation du contenu.

Chapitre 02

Segmentation et structuration des vidéos de présentations

L'indexation et la navigation des vidéos éducatives nécessitent de passer en premier temps par l'organisation structurelle de leurs contenus. Il s'agit de la segmentation temporelle ou encore dite détection des points de transitions pour l'identification de différentes parties disjointes.

Chaque partie ou segment constitue une unité significative, regroupant un ensemble de trames d'un contenu similaire et cohérent vis-à-vis les autres unités, peut être représenté par une seule trame clé. En regroupant les trames clés, une abstraction sémantique est obtenue et peut être utilisée pour l'accès et la navigation.

Dans ce chapitre, nous présenterons la notion de segment et point de transition pour les vidéos de présentations, ainsi que le processus de détection de points de transitions avec ses différentes étapes.

2. Segmentation et structuration des vidéos de présentations

2.1 La segmentation des vidéos de présentations

2.1.1. Définitions

L'indexation, l'annotation et la navigation de tout document repose sur une bonne description de son contenu [27]. Pour décrire au mieux le contenu, il est impératif de passer en premier temps par son organisation. Les systèmes existants exploitent une version du document plus descriptive, plus organisée et moins courte que le document initial. On parle ainsi de l*'abstraction*.

L*'abstraction* : est un processus permettant de trouver un ensemble de trames et générer une vidéo plus courte fournissant à l'utilisateur une abstraction sémantique de toute la vidéo [44].

L'ensemble de trames obtenus est un ensemble de *trames clefs* (*key frames*). Chaque trame représente une information significative et distincte de point de vue utilisateur. Le contenu des vidéos est organisé sous formes de segments contenants des trames de contenu similaire et est représenté par des trames clés via un processus de *segmentation*.

La segmentation : est le processus par lequel une vidéo est partitionnée en un ensemble de segments. Le partitionnement peut être effectué à base de détection de changement de scènes ou détection de segments [4].

L'identification des différents segments se base sur la détection des limites entre eux. Chaque limite est caractérisée par le changement de contenu des trames marqué par un point de transition. La détection des points de transition est une des techniques les plus utilisées pour l'abstraction. Elle est utilisée surtout pour les documents ayant une structure temporelle bien définie [45].

2.2 Organisation structurelle d'une vidéo

L'organisation du contenu consiste à segmenter la séquence vidéo en un ensemble de segments constituant par la suite des unités de traitements. Ceci permet la navigation du contenu qui peut être facilement exploité par l'être humain [46] et aussi pour les applications d'indexation.

On distingue deux types de segmentation : *temporelle* et *spatiale*. La vidéo est découpée en plusieurs parties à plusieurs niveaux. Chaque partie de tout niveau particulier constitue une unité cohérente et distincte des autres unités du même niveau.

La *segmentation temporelle* est la segmentation du contenu en plusieurs parties, chacune constitue une unité sémantique dans le temps. Il s'agit d'une structure hiérarchique qui vise à regrouper au fur et à mesure les entités de chaque niveau ayant une relation sémantique vers un niveau plus haut. A chaque niveau, à l'intérieur d'une entité, les composants ont des contenus similaires et entre différentes entités les contenus sont disjoints.

Un *Segment (video shot)* : est une séquence de trames consécutives interreliées et capturées par une seule caméra entre l'action et l'arrêt de la caméra [47]. Le niveau segment (shot) est considéré le plus approprié pour l'indexation et la navigation. [45]

Une *Scène* peut être composée d'un ou plusieurs segments qui décrivent une unité d'histoire dans une vidéo [47] ayant une même unité sémantique.

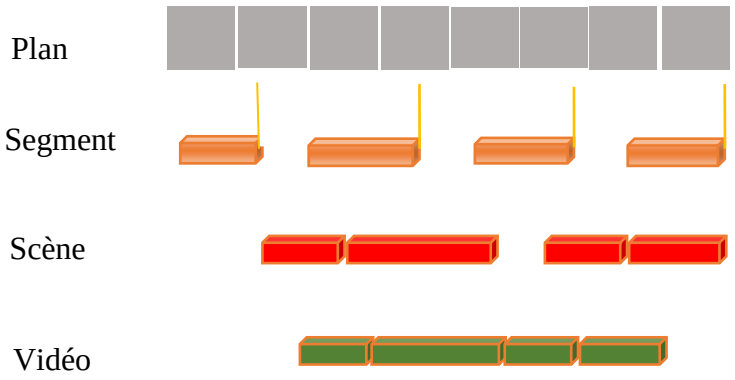

Fig. 2.1. Organisation structurelle hiérarchique du contenu d'une vidéo. [48]

La *segmentation spatiale* est faite à partir des trames clés extraites par segmentation temporelle. Chaque trame représentant une unité structurelle est segmentée en objets distincts.

2.2.1. Organisation structurelle des vidéos de présentations

Les vidéos des présentations sont de nature longue et non structurée. Leurs contenus ne respectent pas la structure hiérarchique dite ordinaire qui se base sur action/arrêt de la caméra. En effet, toute la vidéo est enregistrée entre une seule action/arrêt d'où l'organisation n'est

pas respectée et ne peut être exploitée, d'où la caractéristique *non structurée* est attribuée à ce genre de vidéos.

D'autre part, le contenu sémantique n'est pas relié à l'action de la caméra mais à ce qui est capturé par la caméra. Le cours organisé sous forme d'une présentation permet d'organiser le contenu vidéo en un ensemble de parties selon le topique de la diapositive en cours (Fig. 2.2).

L'absence d'une organisation impose que pour parcourir le contenu ou chercher une information particulière dans une vidéo, la vidéo doit être vue et revue plusieurs fois en avant et en arrière, ce qui est inutile et couteux. Face à ces conditions, une *abstraction sémantique* doit être dérivée afin de permettre l'accès et l'exploitation efficace.

L'exploitation des vidéos éducatives doit faire face aux deux caractéristiques : Longue et Non structurée. Ces conditions doivent être respectées par toutes les méthodes d'abstraction. Il est impératif de passer, en premier temps, par l'organisation de leurs contenus.

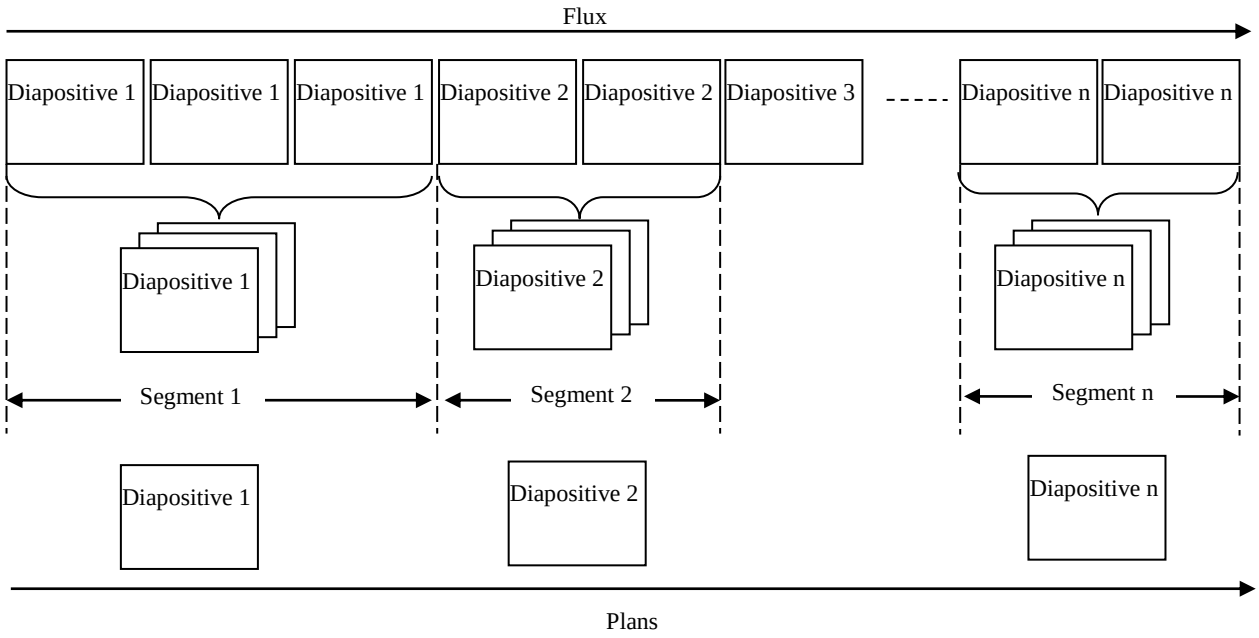

Fig. 2.2. Structure de vidéos de présentations

2.3 Détection de points de transitions (Shot boundary detection)

2.3.1. Segment diapositive

Pour permettre la navigation automatique, les contenus doivent être structurés en un ensemble de segments disjoints appelés *Segments diapositives.* Ils sont considérés comme l'unité fondamentale structurelle pour l'indexation et la navigation par le contenu des vidéos.

La vidéo est segmentée en un ensemble de segments. *Un Segment diapositive (Slide shot) est un ensemble de trames d'une vidéo qui contient la même diapositive. Les limites du segment sont marquées par la transition des diapositives.* [7]

Pour les vidéos de présentations, les segments sont regroupés selon le contenu des diapositives. Les trames contenant la même diapositive sont regroupées en segments séparés par des points de transition qui apparaissent sous forme de changement de diapositive. Ce regroupement reflète au mieux le contenu sémantique de haut niveau, car du point de vue utilisateur, les trames d'une même diapositive correspondent au même topique.

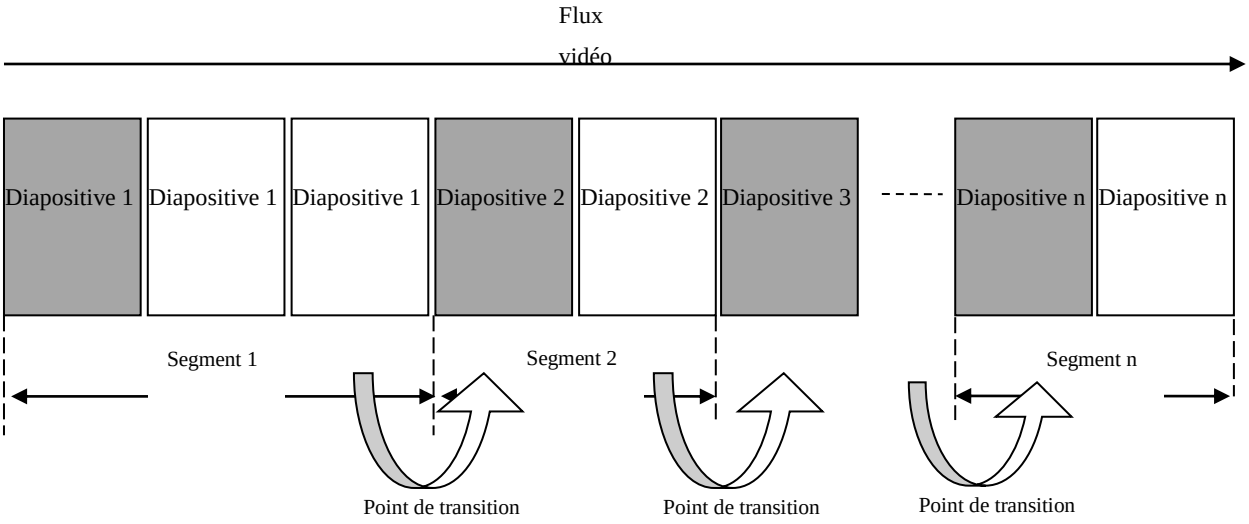

Fig. 2.3. Exemple de segment diapositive et points de transitions

2.3.2. Transitions

2.3.2.1. Définitions

Le point de transition est le point limite entre deux segments adjacents [38].

La détection des limites entres segments adjacents revient à chercher une similarité ou une distance entre les trames adjacentes. Les trames d'un même segment sont similaires,

contrairement à ceux appartenant à des segments différents. Les trames sont dites similaires si la distance entre elles est inférieure à un seuil donné. Ainsi une grande distance revient à un changement du contenu qui marque la fin d'un segment et le début d'un autre.

Pour les vidéos de présentations : un point de transition est le point de changement de diapositive en cours vers une autre.

La détection des transitions dans les vidéos éducatives est une tâche délicate. La plupart des méthodes existantes utilisent une mesure pour comparer la distance entre les caractéristiques visuelles des trames. Cette mesure n'est pas pratique dans le cas de vidéos de présentations où le changement entre différentes diapositives ne provoque pas un changement significatif entre trames adjacents. En effet, les diapositives d'une même présentation ont souvent le même thème, ainsi, le contenu visuel entre les différentes trames est très similaire [49]. Il est clair que les méthodes traditionnelles ne peuvent être appliquées directement à ce genre de vidéos.

Les conditions d'acquisitions et le mouvement de la caméra qui caractérise les vidéos éducatives doivent être aussi prisent en considération. Il est important de trouver un bon compromis entre l'efficacité de description et la détection, pour éviter les faux positives en détection causés par les conditions d'acquisitions.

2.3.2.2. Classification des transitions

Les transitions entre les segments adjacents peuvent se faire de manière brute ou graduelle selon le montage de la vidéo [48]. On distingue ainsi deux types de transitions : brute (CUT) ou graduelle (Gradual Transition GT).

Une transition est dite *brute* si les limites entre deux segments adjacents sont bien définies de manière discrète. Dans le cas de vidéos de présentations, il s'agit d'un changement d'une diapositive de la présentation à une autre. La plupart des vidéos de ce genre présentent cette forme de transitions.

Une transition est dite *graduelle* si aucune limite discrète ne peut être définie entre deux segments adjacents. Le passage d'un segment à un autre se fait de manière progressive. C'est le cas des diapositives contenant des animations. La détection de ce type de transition reste floue et difficile.

2.4 La détection des points de transitions

L'identification des limites entre segments, a pour objectif de déterminer la structuration temporelle des documents par la détection des limites entre les segments.

Le processus de segmentation temporelle est constitué de trois étapes : L'extraction des caractéristiques, calcul de similarité et la détection de points de transitions.

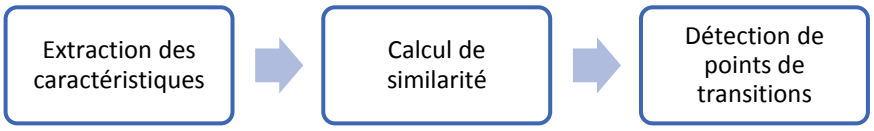

Fig. 2.4. Processus de segmentation des vidéos de présentations

Les systèmes de détection de points de transitions (DPT) identifient les limites des segments en se basant sur le contenu visuel des trames [45]. Ceci peut être justifié par les avantages suivants :

- Le contenu visuel contient la majorité des informations relatives au contenu.
- Les manipulations se font au bas niveau : caractéristiques faciles à extraire et à comparer.

2.4.1. Description du contenu visuel

Pour détecter le changement entre différentes trames, leurs contenus visuels doivent être comparés. Les résultats des comparaisons sont fortement liés aux méthodes de description du contenu. Le choix d'une méthode d'extraction des caractéristiques adéquate et la sélection des caractéristiques appropriés est une étape critique qui doit être établit soigneusement.

Plusieurs méthodes ont été utilisées :

2.4.1.1. Histogramme

L'histogramme est calculé pour chaque trame. Une similarité est calculée entre chaque pair de trames et comparée à un seuil prédéfini [50]. Cette méthode est simple mais elle est utilisée pour le cas où la position de la caméra et celle du présentateur sont fixes. En plus, elle n'est pas assez expressive [30], ce qui ne permet pas une bonne discrimination entre trames des segments différents surtout le cas où les segments appartiennent à la même scène.

2.4.1.2. Contours

Les méthodes d'extraction de contour tel que CANNY [51] est utilisée pour la détection des contours à partir des trames vidéo pour surmonter le problème de luminance. Néanmoins, les performances des contours restent moins effectives que le simple histogramme [45]. En plus, qu'il est connu que leur calcul est difficile et nécessite plus de temps de calcul.

2.4.1.3. Mouvements

Pour faire face au problème de mouvement de la caméra et de l'interlocuteur, les mouvements (vector motions) ont été exploités [52, 53]. Une classification sémantique des trames vidéo dans des catégories est appliquée, tel que le zoom avant, le zoom arrière.

2.4.2. Mesures de similarités

2.4.2.1. Métriques de mesures de similarités

Pour la détection, la distance euclidienne et l'intersection d'histogramme sont les métriques les plus utilisées [45], connues pour leurs simplicité et efficacité [54]. Les descripteurs calculés à partir des trames sont utilisés pour la mesure de distance. Il est clair que la pertinence des résultats obtenus est fortement liée à la mesure utilisée. Ainsi, le choix de la mesure doit respecter aussi la caractéristique sélectionnée.

2.4.2.2. Méthodes de calcul de similarités

2.4.2.2.1. Plans consécutifs

La similarité est calculée et la distance est mesurée entre toute paire de trames adjacents. Cette méthode de calcul prend en considération toutes les trames malgré qu'elle est gourmande en temps de calcul et sensible au bruit. Le bruit produit par les conditions d'acquisition cause une grande distance entre trames adjacentes.

2.4.2.2.2. Fenêtre glissante : N-Plans

La similarité est calculée entre deux ensembles de *k* trames successifs au lieu de deux trames adjacentes. Cette méthode permet d'incorporer des informations contextuelles pour mesurer la similarité, ce qui produit des résultats meilleurs par rapport à la méthode précédente ; mais nécessite plus de temps de calcul.

2.5 Détection et classification des points de transition

Un point de transition est détecté si une distance entre les trames est importante. La grande distance représente la non-similarité entre deux trames et par conséquent avoir du contenu non similaire.

La similarité ou distance (non-similarité) est mesurée par rapport à un critère, qui varie entre l'utilisation d'un seuil ou en se basant sur des méthodes statistiques. Les méthodes utilisant un seuil utilisent une mesure (valeur) calculée en fonction du contenu visuel. Un point de transition est détecté si la distance est supérieure à cette mesure, sinon les trames sont dites similaires et par conséquent appartenait à un même segment. Les méthodes statistiques utilisent des fonctions statistiques tel que la moyenne [22] ou des classificateurs tel que SVM et K-moyens pour l'extraction des trames de transitions.

2.5.1. Méthodes à base d'un seuil

Un seuil est défini pour l'identification des points de transitions. Une transition est détectée si la mesure de similarité calculée est supérieure à ce seuil. Si par contre la valeur calculée est plus petite que le seuil, les trames sont considérées similaires, ainsi, l'absence d'une transition.

Le calcul du seuil peut se faire d'une manière globale ou adaptative.

2.5.1.1. Calcul global

Un seuil global est calculé une seule fois en fonction des informations issues du contenu [30]. Il s'agit d'une valeur unique utilisée pour toutes les mesures de comparaison le long de la vidéo.

Cette méthode de calcul est la plus simple. Néanmoins, elle ne reporte pas les changements locaux du contenu, ce qui diminue l'efficacité de la détection à cause du fait que le seuil est calculé une seule fois de manière générale pour toute la vidéo.

2.5.1.2. Calcul adaptatif

Reviens à calculer le seuil localement pour un ensemble de trames au lieu de le calculer pour toute la vidéo. Il est adopté par plusieurs méthodes [7, 30]. Le seuil est estimé pour chaque séquence de trames, une fenêtre glissante est utilisée pour regrouper un nombre limité

de trames et qui varie le long de la vidéo. Ceci permet de mieux prendre en considération les changements locaux du contenu à l'intérieur de la fenêtre de calcul. Néanmoins, le calcul est plus difficile par rapport au calcul global. Plusieurs paramètres doivent être configurés pour avoir un meilleur résultat tel que la définition de la taille de la fenêtre, qui nécessite une bonne connaissance des caractéristiques de la vidéo.

2.5.1.3. Combinaison des méthodes globales et adaptatives

Une solution plus efficace peut être produite en combinant les avantages des deux méthodes de calcul, si le seuil adaptif est estimé à base du seuil global. Mais la relation entre les deux est difficiles à définir.

2.5.2. Calcul a base des méthodes statistiques

Les trames d'une vidéo peuvent êtres catégorisés en deux ensembles. Un ensemble qui regroupe les points de transitions (non similaires) et l'autre regroupe les trames similaires. Ainsi, les méthodes de classification peuvent être utilisées. Selon le type de classifieur utilisé, nécessitant un apprentissage ou non, on distingue deux types d'approches : celles dites approches supervisées et d'autres dites approches non supervisées.

2.5.2.1. Approches supervisées

Les méthodes de cette approche utilisent un classifieur supervisé, tel que le SVM et Adaboost [55]. Ces approches [22, 56] utilisent des classifieurs trainés pour regrouper les trames qui constituent les points de transitions. Mais, le temps d'apprentissage et le choix de l'ensemble d'apprentissage adéquat reste une limitation majeure. En plus, la sélection de la bonne caractéristique et sa bonne représentation, pour avoir les résultats recherchés, est encore un problème non résolu de la classification supervisée.

Trouvez le bon ensemble d'apprentissage n'est pas évident, surtout dans le cas des vidéos de présentations. En effet, il existe plusieurs types et thèmes de présentations, aussi avec différents couleurs, styles et fonts. Ainsi, tout ensemble choisi ne peut en aucun cas couvrir toutes les caractéristiques existantes.

2.5.2.2. Approches non supervisées

Ces approches utilisent un classificateur non supervisé tel que K-moyen ou Fuzzy K-moyen pour classifier les trames en points de transitions et autre dites similaires. Ces

approches ne nécessitent pas une étape d'apprentissage, mais restent gourmandent en temps de calcul.

2.6 Approches existantes

La segmentation des vidéos est un domaine de recherche actif et plusieurs études ont étaient élaborées [4, 44]. Les solutions proposées sont souvent dédiées à des vidéos dites ordinaires. De nos jours, plus d'intérêt est attribué aux vidéos éducatives. Dans ce qui suit, nous présenterons les différentes méthodes d'analyse structurelle pour ce genre particulier de vidéos.

Une des premières méthodes de segmentation des vidéos de présentations est celle proposée utilisant une caméra fixe. Mukhopadhyay et al. [50] calculent l'histogramme à partir d'une trame, après, ils comparent les régions des trames adjacentes. Dans le cas où le mouvement de la caméra est permis (configuration non stationnaire est adoptée), d'autres méthodes à base de contours [5] ou de mouvement ont été proposées qui sont moins sensibles à la luminance, mais plus gourmandes en temps de calcul et plus complexes. D'autres méthodes [52, 57] se basent sur la détection des mouvements, car avec la présence d'un mouvement il n'y aura pas de changement de diapositives, exemple le cas du zoom, mais ce n'est pas le cas pour toutes les vidéos.

Dans [12, 13, 30, 51] on utilise des données supplémentaires appelées métadonnées. Mais la présence des métadonnées supplémentaires n'est pas toujours possible.

Les caractéristiques extraites sont utilisées pour les mesures de similarité entre les trames adjacentes par paires [7, 58]. Plusieurs métriques ont été utilisées telles que la distance euclidienne et l'intersection d'histogrammes. Ces méthodes sont gourmandes en temps de calcul et effectuent des comparaisons inutiles, du fait que la similarité est calculée pour chaque paire de trames successives. Les méthodes récentes [49, 51] calculent la similarité entre les trames dans une fenêtre coulissante afin d'éviter les comparaisons inutiles et d'incorporer plus d'informations contextuelles. Cependant, ils considèrent que la configuration est fixe (caméra fixe).

L'identification des points de transitions à base d'un seuil est la méthode la plus utilisée. Jeong et al. [7] identifient les points de transitions à base d'un seuil global. Il est calculé à partir de la moyenne et la déviation d'un ensemble de trames. Une autre méthode utilise un

seuil adaptatif [5] basé sur des composants connectés C.C. où le nombre de C.C est limité pour un seuil. Cependant, cette méthode considère le format multi-scènes qui n'est pas le cas de la plupart de vidéos. D'autres méthodes incorporent plus d'informations contextuelles par l'utilisation d'une fenêtre glissante.

2.7 Discussions

Les méthodes proposées présentent les limitations suivantes :

- Utilisent des données supplémentaires qui ne sont pas toujours disponibles.
- La configuration fixe qui n'est pas le cas de la plupart des vidéos.
- Solutions avec des conditions spécifiques.

Les méthodes de détection de points de transitions peuvent être regroupées en deux catégories : celles qui adoptent une configuration fixe et celles qui utilisent des configurations non fixes. Les méthodes de la première catégorie sont devenues rapidement inadéquates, du fait que la plupart des vidéos récentes utilisent une configuration non fixe. Ces vidéos sont enregistrées :

- par des caméras mobiles ordinaires.
- par des gens non professionnels.
- L'interlocuteur est libre à se déplacé.

Les conditions d'acquisition ont une grande influence sur la qualité de la vidéo. Ces conditions affectent le contenu visuel des trames et par conséquent affectent les performances de description et de segmentation du contenu. Les solutions existantes présentent de bonnes performances vis-à-vis des conditions d'acquisition spécifiques et bien définies. Mais les grandes variétés en luminance et distorsion, dû au mouvement, les rendent vite inadéquates sous d'autres conditions.

Il est important de prendre ces contraintes en considération et de trouver une caractéristique qui soit à la fois expressive, pour permettre de capturer au mieux le changement du contenu dans tous les trames, mais assez invariante aux différents bruits produits par les conditions d'acquisition.

En conclusion, Toute méthode proposée doit répondre aux exigences suivantes :

- La nature des vidéos : longues et non structurées.

- La basse qualité due aux conditions d'acquisition : doit couvrir le maximum de conditions.
- Le problème de luminance qui mène à de fausses détections.
- Invariance au mouvement de la caméra qui provoque un grand changement dans le contenu visuel des trames, qui n'est pas relié à son contenu sémantique.
- Doit être indépendante de toute information supplémentaire.

2.8 Conclusion

La segmentation et la structuration du contenu des documents multimédias est une étape importante de tout système d'indexation par le contenu. Plusieurs méthodes ont été proposées mais le problème reste toujours loin d'être résolu, surtout pour ce type particulier de vidéos : *vidéos de présentations* qui présentent plusieurs contraintes. Les solutions proposées pour les vidéos éducatives soufrent d'êtres spécifiques à des conditions prédéfinies et par conséquent l'absence d'une solution générale.

L'étape de segmentation a pour objectif d'organiser le contenu et de fournir une abstraction qui peut être par la suite utilisées pour l'extraction des caractéristiques pour l'indexation. Dans le chapitre suivant nous présenterons les données de haut niveau utilisées pour l'indexation et plus précisément pour l'extraction des caractéristiques visuelles qui constituent l'élément le plus important dans le processus d'indexation.

Chapitre 03

Le texte de scène : Un descripteur sémantique pour l'indexation

Le contenu visuel des images ou des trames vari entre le texte, les objets et la motion humaine. Les recherches ont montré que le texte est le premier objet aperçu par l'être humain [59]. Par conséquent, il est utilisé intuitivement pour la description des informations contextuelles.

L'extraction automatique du texte est une tâche délicate qui consiste généralement en plusieurs étapes : détection et localisation, amélioration, segmentation et enfin la reconnaissance optique des caractères [60]. La difficulté d'extraction est issue des conditions d'acquisition des images/vidéos contenant du texte, qui peuvent êtres de basse qualité (contraste, luminance) ou êtres complexe avec beaucoup de couleurs et des arrières trames texturés. En plus, le texte en lui-même peut varier en font, style, couleur, orientation et alignement.

Dans ce qui suit nous présenterons les systèmes d'extractions des informations textuelles, ainsi que les différentes approches et méthodes utilisées. On passera par une description détaillée de chaque étape du processus ainsi que de différentes solutions existantes.

3. Le texte de scène : Un descripteur sémantique pour l'indexation

3.1 Introduction

De nos jours, la plupart des appareils mobiles incorporent des appareils à photos et permettent une large acquisition de documents numériques. Le texte fait une partie importante du contenu capturé tel que les panneaux routiers, la projection d'un ordinateur, les plaques d'immatriculation des véhicules, les pages des livres capturés comme images. En effet, le texte peut décrire le contenu et donne des indications sur le contexte. Ce texte peut être utilisé pour l'indexation et l'annotation automatique à base de contenu.

3.2 Le texte dans les images et vidéos

Les informations textuelles dans les images à scène naturelle font parties des objets inclus dans la scène capturée lors de l'acquisition. L'exploitation des systèmes de reconnaissance optique de caractères traditionnels ne manipulent que du texte imprimé sur un arrière-trame généralement blanc. Le passage par un ensemble de méthodes pour préparer les images à scène naturelle et les faire rapprocher le plus à ceux des documents scannés est indispensable. Ces méthodes couvrent la détection et l'extraction des informations textuelles.

3.2.1. Importance du texte

L'importance du texte pour la description du contenu sémantique des images/trames est due essentiellement aux facteurs suivants :

- Le texte est le premier objet détecté par l'œil humain [59].
- L'efficacité pour la description du contenu contextuel. [60] Ainsi, même une petite quantité peut apporter des informations importantes sur le contenu des images tel que les noms des rues utilisés dans les systèmes de navigations.
- Facile à extraire par rapport aux autres éléments sémantiques [60] et exploité facilement par l'être humain et la machine.
- Large domaine d'exploitation.

3.2.2. Types de texte

Les informations textuelles dans les images/trames vidéo peuvent être du texte super imposé ou du texte de scène dans les images.

3.2.2.1. Texte de légende (super imposé)

C'est du texte ajouté artificiellement sur les images/trames dans une étape dite d'édition. Le texte est super imposé afin de décrire le contenu de l'image ou la vidéo. Des exemples du texte super imposé : le sous-titrage dans les informations télévisées, le score dans les vidéos sportives. Le fait que le texte est ajouté artificiellement aux images/trames, il possède des caractéristiques spécifiques tel que l'alignement horizontal, l'endroit d'insertion, le mouvement linéaire, la luminance indépendante de l'images/trame, en plus de la taille, la fonte et l'espace inter/intra caractères qui est le même dans toute la zone du texte.

(a)

(b)

Fig. 3.1. Exemples des images contenant du texte. (a) texte légende ; (b) texte de scène. [59]

3.2.2.2. Texte de scène (enfoui)

Le texte de scène apparait naturellement dans les images capturées par la caméra lors de l'acquisition. Il est plus difficile à détecter le texte du fait qu'il peut être dans n'importe quelle orientation, avec différentes couleurs et tailles. L'autre difficulté réside dans les conditions d'acquisition des images/vidéos ou l'orientation de la caméra, l'illumination, le mouvement qui peuvent avoir un grand impact indésirable sur le texte. Comme exemple du texte de scène, les noms des joueurs dans les vidéos sportives, les noms de rues, les indications sur les panneaux routiers.

Le texte de scène est une région qui fait partie de l'image même. Ce texte est souvent lié au contenu sémantique, ce qui motive son utilisation comme un descripteur pour l'indexation par le contenu.

Les documents numériques peuvent aussi être capturés par la caméra tel que les pages ou couvertures des livres. Il s'agit de l'utilisation d'un scanner. Ces images, contenants généralement du texte sur un arrière-trame uniforme ou coloré, sont considérées comme des images de texte de scène.

3.2.3. Caractéristiques du texte

Le texte apparent sur une image ou une séquence vidéo possède plusieurs caractéristiques : des caractéristiques géométriques, la couleur, la fonte, le mouvement, le contour, la déformation de perspective et la compression.

- **Les caractéristiques géométriques :** Il s'agit de la taille, l'alignement et l'espace inter-caractères. La *taille* du texte varie dans le même mot, même phrase, même image et même d'une application à une autre. *L'espace inter-caractères* est le même (Uniform) entre les caractères d'un seul mot. Cette caractéristique est importante lors du regroupement des caractères individuels en mots. *L'alignement* du texte dépend de la nature du texte, par exemple dans une légende il est souvent aligné horizontalement. Dans le cas du texte de scène, il peut être aligné dans n'importe quelle direction, ce qui provoque une distorsion de sa forme géométrique. Ce qui complique la tâche des systèmes d'extraction d'informations textuelles.
- **La couleur :** Généralement pour une meilleure visualisation, le texte possède une couleur différente de l'arrière-trame. Il peut avoir différentes couleurs, soit dans une chaîne de caractères (polychrome) ou dans le même mot (monochrome) [60].
- **Mouvement :** Dans une vidéo, le même texte peut figurer dans plusieurs trames successives, ce qui permet son suivi et la possibilité de correction et d'amélioration. Le texte de scène peut avoir un mouvement aléatoire contrairement au texte super imposé qui a un mouvement linéaire [60].
- **Contour :** Le texte possède généralement de forts contours aux limites des caractères avec l'arrière-trame.
- **Déformation de perspective :** Provoquée par l'angle de la caméra lors de l'acquisition (camera non frontale). Cette déformation influe considérablement sur les performances du système d'extraction des informations textuelles TIE [59, 61].

- **Compression :** Un grand nombre d'images/vidéos sont manipulées dans un format compressé. Par conséquent, il sera utile de les utiliser sans avoir à les décompresser [60].

3.2.4. Difficultés que présente le texte

D'après les définitions et caractéristiques, il est clair que le texte super imposé est plus facile à extraire que le texte de scène. En effet, pour le texte super imposé, plusieurs contraintes sont imposées tel que : l'alignement horizontal, l'insertion à des endroits spécifiques et l'espace inter et intra caractères est le même dans toute la zone du texte ce qui facilite la tâche d'extraction. Dans ce qui suit, nous citons les différentes difficultés présentées par le texte de scène que doit surmonter un système d'extraction d'information textuelle :

- Absence d'information préalable sur la présence ou l'absence du texte. Dans le cas de présence la localisation dans les images/trames et aussi inconnue.
- Une partie du texte peut être occlue [62].
- Les variétés dans les différentes propriétés : couleurs, fontes, taille et alignement dans le même mot ou dans un ensemble de mots.
- Les conditions d'acquisition tel que : l'angle de d'acquisition lors de la prise des photos ou vidéos provoque une déformation de perspective, variation de luminance, l'arrière-trame complexe, l'ombre et le mouvement.
- Les images/vidéos sont généralement de faible résolution.

3.2.5. Applications

Les avantages du texte cités précédemment motivent son utilisation dans plusieurs domaines, parmi lesquels :

- *Indexation par le contenu sémantique des vidéos* : qui peut être utilisée pour l'abstraction et l'annotation vidéo, détection de logo TV. Le but est d'indexer et manipuler de manière efficace de grandes bases de données multimédias selon leurs contenus sémantiques où le texte extrait de manière automatique est utilisé comme un descripteur de haut niveau ce qui améliore considérablement l'efficacité de ces systèmes.
- *Analyse et indexation des vidéos de sport :* c'est l'une des applications importantes, développée pour le grand nombre des amateurs des nombreux sports existants. Ces systèmes manipulent une large quantité de vidéos qui nécessitent le développement de plusieurs applications permettant l'annotation, la reconnaissance d'événement de tir ou encore l'identification automatique et suivi des joueurs [63].

- *La publicité dans les vidéos* : Les technologies récentes offrent diverses possibilités permettant de fournir la diffusion des vidéos, qui est devenue de plus en plus fameuse et par conséquent, un domaine intéressant pour les activités commerciales tel que le marketing. La publicité est ajoutée selon le contenu textuelle ou sémantique [64] ; ainsi, différents utilisateurs sont visés avec plus de précision selon leurs domaines d'intérêt. YouTube, AOL Video et Yahoo! Video ajoute des publicités relatives au contenu via une analyse textuelle ou suite à une requête. [59].
- *Reconnaissance des caractères* sans avoir passé par un scanner comme les documents historiques et les pages des grands livres [61]. Les pages des documents sont capturées par des appareils photos, ceci facilite et permet de conserver les documents importants tels que les documents historiques.
- *Reconnaissance de plaques d'immatriculation :* Les numéros des plaques d'immatriculation des véhicules sont exploités par plusieurs applications comme la sécurité et le contrôle d'accès, le parking, le contrôle de circulation. Ces systèmes se composent généralement de trois étapes principales : 1) extraction de la région de la plaque 2) segmentation des caractères 3) reconnaissance de ces derniers [65]. Les systèmes de reconnaissance de plaques d'immatriculation doivent faire face aux problèmes liés aux conditions d'acquisition et la luminance non-uniforme ; et ne marchent pas que dans des conditions limitées et spécifiques.
- *Les systèmes de navigation :* Ces systèmes exploitent les informations des panneaux routiers et d'indication. Les systèmes de navigation routières sont un exemple basé sur le texte extrait des panneaux routiers, afin de diminuer les risques, Ces systèmes fournissent des informations préalables sur l'état de la route et la circulation pour éviter les accidents (eg. [66]). Un autre exemple, les systèmes d'aides des personnes à vision limitée ou aveugle [10] qui permettent d'extraire le texte à partir des objets ou panneaux pour aider ces personnes à éviter les obstacles ou trouver leurs chemins sans assistance. Les robots mobiles aussi utilisent les informations textuelles incluses dans les panneaux ou les sites d'intérêts (Landmark) pour leur navigation ; ainsi, le texte de scène est utilisé pour décrire le contexte qui sera utile pour la planification de trajet ou la navigation par objectifs [67]. Les systèmes de navigation doivent avoir un haut taux de précision, un faible taux de faux positifs et répondent en temps réel [59].
- *Recherche d'images par le contenu :* où l'image est recherchée par des mots clés reliées au contenu sémantique, ces mots clés sont extraient de ces images.

- *Dictionnaire mobile :* Pour une traduction instantanée via des appareils mobiles.

3.3 Architecture générale des systèmes d'extraction de texte

Le texte de scène possède plusieurs caractéristiques susceptibles à plusieurs variations :

- L'absence des informations préalables sur l'existence ou non du texte ;
- Position, direction, couleur, alignement ;
- L'arrière-plan complexe ;
- La variation de la luminance.

Et par conséquent, la tâche de l'extraction est complexe et délicate. Ce qui émerge le développement et l'exploitation de techniques spécifiques pour surmonter ces difficultés. D'autre part, la reconnaissance du texte de scène est aussi affectée par les destructions géométriques, la luminance non-uniforme, la faible résolution et même la décoration du texte [68].

Le processus d'extraction d'informations textuelles est un processus complexe composé de cinq étapes [60] : la détection, la localisation et le suivi, l'extraction, l'amélioration et enfin la reconnaissance. La détection a comme objectif de définir la présence ou non du texte qui sera par la suite localisé sous forme de régions qui peuvent être suivies dans des trames successives dans le cas d'une séquence vidéo. L'extraction est effectuée par une segmentation du texte qui sera nettoyé par des techniques d'amélioration afin d'avoir une image binaire prête à être utilisée, par la suite, par la reconnaissance. Le schéma suivant décrit le processus cité précédemment.

L'acquisition des images/vidéos peut être vue comme une étape additionnelle à ajouter au processus d'extraction d'informations textuelles. En effet, Uchida [69] définit ce processus comme étant formé de trois étapes : acquisition, localisation et reconnaissance. Qui n'est qu'un regroupement des étapes décrites précédemment en plus de l'étape d'acquisition. Il est vital de prendre cette étape en considération puisque les résultats des étapes qui suivent sont fortement liés aux conditions d'acquisition. Ainsi, l'étape d'acquisition et prétraitement doit être ajoutée au système TIE et par conséquent le développement des techniques spécifiques. Dans ce qui suit, les différentes étapes sont présentées avec plus de détails.

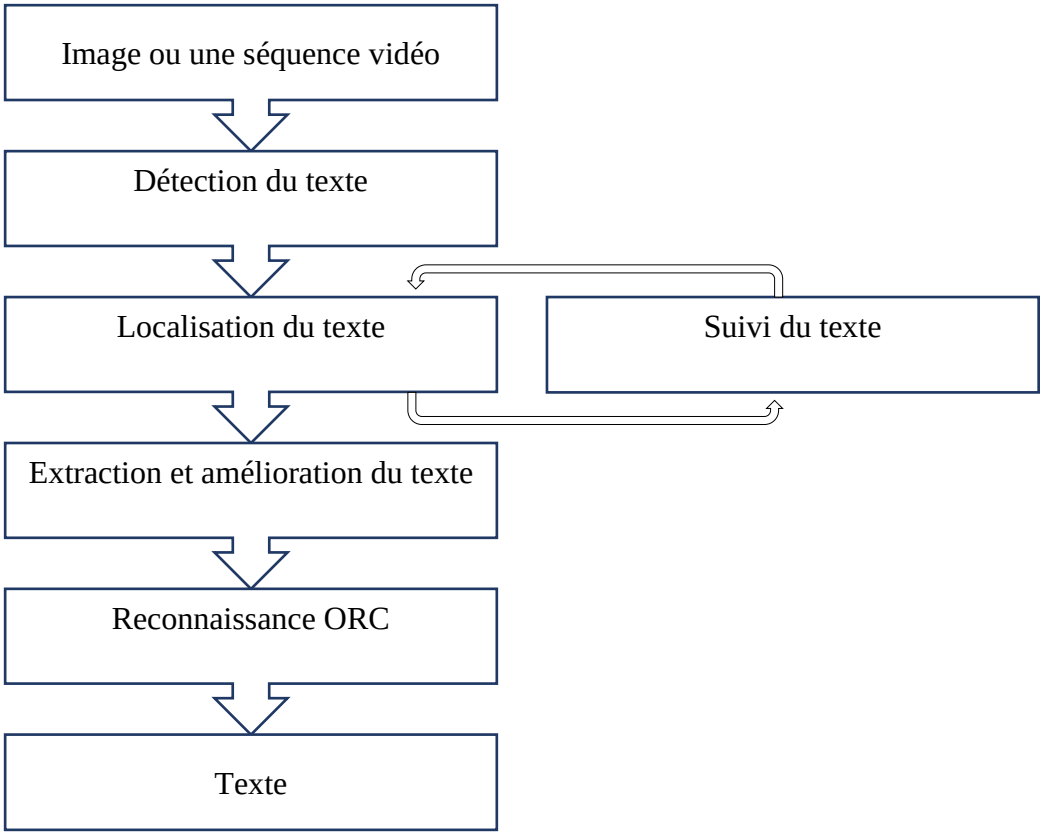

Fig. 3.2. Architecture générale d'un système d'extraction d'informations textuelles EIT. [59]

3.3.1. Acquisition et prétraitement

Selon les conditions d'acquisition, les images/vidéos peuvent êtres de faible résolution, floue ou subies à des distorsions de perspective, luminance non-uniforme ou même un arrière-trame texturé complexe. Combiner ces conditions avec les différentes variations de police, couleur, alignement, taille et orientation du texte rend la tâche d'extraction d'informations textuelles très difficile. Ceci impose l'utilisation des techniques de prétraitement pour faciliter la tâche des étapes suivantes et, par conséquent, améliorer les résultats du système en général.

Le prétraitement pour la détection du texte désigne l'utilisation d'un ensemble d'opérations de bas niveau à fin d'améliorer la qualité des images/trames en éliminant les dégradations indésirables et améliorer les caractéristiques importantes [59]. Le processus de prétraitement est généralement composé de quatre étapes :1) segmentation des régions de l'image en régions contenant probablement du texte et régions ne comportant pas du texte ; 2) filtrage des régions très grandes ou très petites dont il est impossible qu'ils contiennent du texte ; 3) analyse du mouvement en suivant la position du texte ; 4) amélioration de la qualité

de l'image [59, 60]. Il est clair que les quatre étapes citées précédemment ne sont pas obligatoires, l'une ou l'autre sont utilisées selon les spécificités des images et du domaine d'application.

Dans le cas des systèmes de détection de texte, les techniques de prétraitement appliquent des opérateurs aux images/trames en vue de séparer, de manière efficace, les régions qui contiennent probablement du texte de celles qui contiennent d'autres objets tel que l'arrière-trame. Ces opérateurs sont aussi exploités pour améliorer la qualité des images/trames. Les opérateurs les plus utilisés sont des opérateurs locaux, tel que les opérateurs morphologiques, traitements à l'aide de voisinages comme le Filtre médian.

a) Prétraitement basé sur la couleur : Les images naturelles peuvent contenir une variété des objets et par conséquent dans une même images/trames. Différentes couleurs sont influencées par les conditions d'acquisition comme la luminance qui rend la tâche de détection et localisation plus délicate. Des techniques basées sur la couleur, tel que la transformation entre les espaces de couleurs, la réduction de couleurs et la classification basée sur la couleur, sont utilisées fréquemment dans les systèmes de détection de texte.

b) Prétraitement basé sur la texture : Généralement le texte possède une texture différente de son arrière-trame, ce qui permet l'exploitation des techniques d'analyse de texture tel que les filtres de Gabor, les ondelettes et la Différence de Gradient Maximale (MGD) pour l'extraction du texte. Néanmoins, ces techniques soufrent des limitations des techniques d'analyse textuelle tel que la grande complexité de calcul. En plus, ces techniques ne peuvent pas détecter le texte incliné [59].

3.3.1.1. La segmentation

C'est le processus de diviser une image en un ensemble de régions de forte corrélation [70]. Les pixels sont associés à des régions construites à partir des caractéristiques extraites de l'image originale [71]. La segmentation joue un rôle important dans la plupart des applications manipulant des images tel que les systèmes d'extraction des informations textuelles [59]. Le résultat influe par la suite sur les résultats des étapes suivantes. Idéalement, l'objectif est d'avoir des régions contenant du texte bien séparé de l'arrière-trame, ce qui diminue la complexité et limite le traitement des étapes suivantes à des régions restreintes. Les méthodes de segmentation suivent deux stratégies : de bas en haut ou de haut en bas. Dans TIE, les systèmes utilisent la segmentation de bas en haut à cause de l'absence

d'information préalable sur le texte, ainsi, les pixels sont regroupés en régions en se basant sur des caractéristiques de bas niveau.

La segmentation de bas en haut comporte trois classes de méthodes : Seuillage (Thresholding) basé sur les régions et autre sur les contours. Le seuillage d'une image à niveaux de gris consiste à déterminer une valeur *seuil* afin de séparer les objets de leurs arrière-trames. Il s'agit de la forme la plus simple de segmentation, rapide et fiable. Les méthodes basées sur les régions ou sur les contours sont des méthodes exploitant la continuité et la discontinuité des régions. Les méthodes basées sur les contours ont pour but de détecter les frontières des régions en utilisant des opérateurs de détection de contours qui cherchent la discontinuité dans les niveaux de gris, la couleur ou la texture. Les méthodes basées sur les régions utilisent des critères d'homogénéité comme les niveaux de gris, la couleur ou la forme pour former des régions dites homogènes [70].

3.3.1.2. L'analyse du mouvement

Dans le traitement des vidéos, plusieurs trames consécutives peuvent contenir le même texte. Dans le cas de mouvement, le texte change de position par quelques pixels sur chaque trame, ce qui permet son suivi ainsi que d'améliorer et compléter les régions déformées en profitant de la redondance issue des trames différents.

Fig. 3.3. Exemple de l'amélioration qu'apporte l'analyse du mouvement pour la détection du texte : (a) l'image originale, (b) Extraction, (c) Amélioration apportée par l'exploitation de la redondance d'informations sur les trames consécutives [61].

3.3.1.3. Les techniques d'amélioration

L'amélioration peut varier entre : augmenter la résolution, correction de l'angle d'acquisition ou d'éliminer le bruit.

Les images de faible résolution doivent subir des traitements pour améliorer la résolution et par conséquent améliorer sa visibilité [72]. Ces algorithmes ont pour objectif d'obtenir une image à haute résolution (HR) à partir d'une autre de basse résolution (LR). Selon le nombre d'images en entrée, soit une seule ou plusieurs images, on distingue deux classes d'algorithmes de super résolution [72, 73].

L'angle d'acquisition provoque une déformation de perspective ayant un grand impact sur l'étape de reconnaissance. Les deux paramètres (la position et la direction de la caméra par rapport au texte acquit) varient sans limites et par conséquent complique la tâche d'estimation de correction ou d'élimination de la déformation de perspective. Dans le cas des documents capturés par caméra, les contours sont exploités. Les lignes du texte peuvent aussi être utilisées pour estimer la direction du texte dans le cas où les contours ne peuvent être extraient. La forme horizontale des caractères de certaines langues comme l'anglais peut être utilisée, mais les techniques exploitant cette propriété sont gourmandes en temps calcul.

Les images floues sont souvent le résultat d'une mauvaise concentration ou d'un mouvement, on parle d'une convolution de l'image originale par une fonction PSF (Point Spread Function). Cette fonction représente comment se trouve chaque point de l'image originale dans l'image floue, ainsi, l'objectif est d'estimer cette fonction et par la suite récupérer l'image originale [74, 75].

3.3.2. Détection et localisation

La détection et la localisation du texte consiste à déterminer la présence du texte dans une image/trame et trouver sa position exacte.

L'étape de détection et localisation doit faire face aux caractéristiques qu'impose le texte de scène en plus de l'absence de toutes informations sur sa présence/absence ou sa position dans l'image/trame. Plusieurs étapes doivent avoir lieu afin de surmonter ces difficultés. La figure (Fig. 3.4) illustre le processus typique à suivre par les méthodes de détection et localisation. En premier temps l'image/trame est divisée en régions qui seront utilisées pour calculer des caractéristiques qui permettent de déterminer les régions contenantes du texte. Ces dernières sont regroupées dans d'autres régions plus grandes selon les informations de voisinages et entourées par des rectangles.

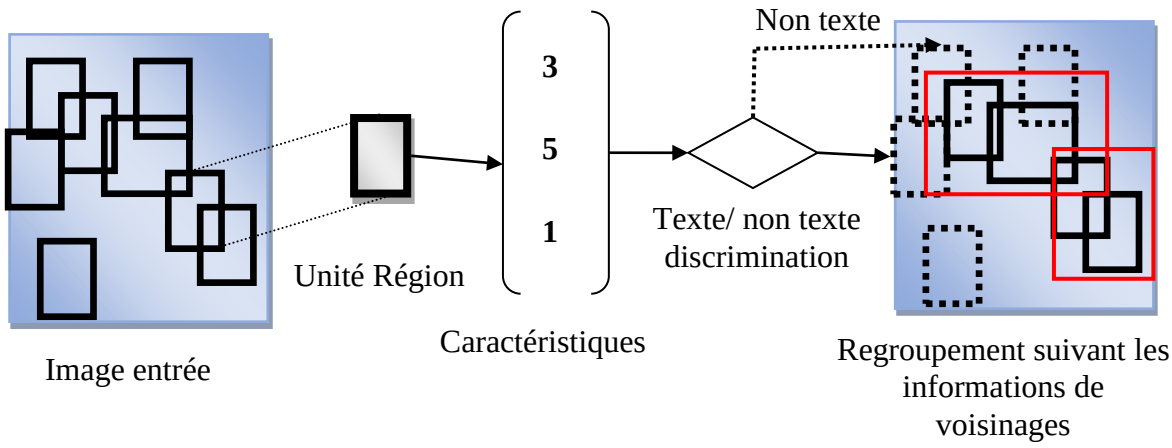

Fig. 3.4. Processus typique de détection et localisation de texte [68].

Les méthodes de détection et localisation de texte peuvent être regroupées en deux grandes classes : méthodes basées sur les régions et d'autres basées sur la texture. La combinaison des techniques de ces deux classes donne une autre classe dite hybride. Vu leurs importances et leurs variétés, les méthodes de détection seront discutées plus en détail dans la section suivante.

3.3.3. Extraction, amélioration et suivi

Les systèmes de reconnaissance optique des caractères (OCR) commerciaux, bien développés pour les documents scannés, produisent de très bons résultats pour les documents comportant du texte sur un arrière-trame uniforme (souvent un texte en noir sur un arrière-trame blanc), ce qui n'est pas toujours le cas du texte de scène qui a souvent un faible contraste, faible résolution et un complexe arrière-trame, en plus des variations de luminance et des déformations de perspective ce qui impose l'utilisation des méthodes d'amélioration et la préparation des images du texte pour l'entrée des systèmes de reconnaissance des caractères.

L'extraction est le processus de segmentation du texte de l'arrière-trame et donne une image binaire composée du texte sur un arrière-trame uniforme. Le but est d'avoir une image binaire (comme celle d'un document scanné) sans perdre les informations textuelles, qui est une tâche difficile avec toutes les contraintes qu'impose les images naturelles.

L'amélioration du texte doit avoir lieu dans le cas où le texte est de faible résolution ou susceptible au bruit. Plusieurs techniques ont été utilisées qui opèrent sur une image ou un ensemble de trames successives. Les techniques discutées dans la section de prétraitement peuvent être utilisées en plus de l'application des méthodes de binarisation.

La binarisation est une opération importante pour l'extraction des informations textuelles, qui permet d'améliorer le résultat de reconnaissance à base des systèmes OCR très performants sur des images des documents scannés binaires. Ainsi il est important de transformer les images de scènes naturelles à autres ayant les mêmes caractéristiques des documents scannés comme avoir un texte sur un arrière-trame blanc.

Le suivi du texte sur plusieurs trames consécutives permet de réduire le temps de réponse de tout le système en gardant la position du texte déjà détecté, et par conséquent, éviter d'appliquer l'étape de localisation à nouveau pour chaque trame qui parait intéressant surtout si le temps de suivi est moins que celui de détection et de localisation.

3.3.4. Reconnaissance

L'étape de reconnaissance consiste à transformer une image binaire à un texte brut. La reconnaissance est faite à l'aide d'un système OCR ou à l'aide des algorithmes de classification. L'existence des systèmes OCR bien développés et bien testés motivent leur large utilisation, mais ils exigent une image binaire en entrée pour fournir de bons résultats qui est une tâche difficile pour les images/trames naturelles. Une autre possibilité est de faire un apprentissage supervisé à l'aide d'exemples de texte. L'utilisation des algorithmes de classification nécessitent un apprentissage à base d'un grand nombre d'exemples ce qui limite leurs champs d'application à un ensemble restreint de données [59].

Généralement la reconnaissance des caractères se fait par des systèmes OCR (Optical Character Recognition). Ces systèmes sont conçus au début pour les documents scannés ayant une bonne résolution. Le texte est tapé sur du papier blanc qui n'est pas le cas du texte de scène qui peut avoir de grandes variétés de couleurs, tailles, fontes en plus de décorations ou déformations causées par les conditions d'acquisition ou engendrées par l'arrière-trame. L'étape d'amélioration vise à préparer le texte segmenté et le rendre le plus proche possible au texte des documents scannés.

3.3.5. Evaluation des performances

Les performances d'un système d'extraction d'informations textuelles peuvent être mesurées par différentes mesures. Les plus utilisées [76] sont le *taux de détection*, la *précision*, le *rappel* et la *F-mesure*.

Le taux de détection est le ratio entre le *nombre de blocs de texte* détectés et le *nombre total de blocs de texte réellement existants* dans l'image/trame.

$$Taux\ de\ détection\ = \frac{nombre\ de\ \text{blocs de}\ texte\ détectés}{Nombre\ total\ de\ blocs\ de\ texte\ existants} \tag{3.1}$$

Le *nombre de blocs de textes* peut être le nombre de rectangles englobants un caractère, un mot complet ou même un ensemble de mots.

La précision *P* est le ratio entre le *nombre de caractères correctement détectés* par rapport au *nombre total de caractères détectés*.

$$P = \frac{Nombre\ de\ cractères\ correctement\ detectés}{Nombre\ total\ de\ caractères\ detectés} \tag{3.2}$$

Rappel *R* est le ratio entre le *nombre de caractères correctement détectés* par rapport au *nombre total de caractères*.

$$R = \frac{Nombre\ de\ cractères\ correctement\ detectés}{Nombre\ total\ de\ caractères} \tag{3.3}$$

La mesure de qualité *F* combine les mesures rappel *R* et précision *P*, généralement pondérées par 0.5 pour les systèmes d'extraction d'informations textuelles qui donne une importance égale aux deux critères *R* et *P*.

$$F = 2 * \frac{R*P}{R+P} \tag{3.4}$$

3.4 Méthodes d'extraction de texte : Revue des méthodes

Les méthodes d'extraction de texte, comme mentionné plus haut, exploitent des caractéristiques extraites de l'image pour déterminer la présence ou non du texte et par la suite sa position. Selon les caractéristiques exploitées on distingue les deux classes de méthodes :

celles basées sur la texture et celles basées sur les régions. La combinaison des méthodes des deux classes génère une troisième classe dite hybride.

La plupart des méthodes de détection et localisation suivent une approche ascendante de bas en haut pour l'identification des régions contenantes du texte. Les pixels sont en premier lieu regroupés en caractères ensuite en mots ou ensemble de mots formant une ligne de texte. Le regroupement des pixels se fait selon des caractéristiques comme la couleur, les niveaux de gris ou la largeur constante du stroke (épaisseur de trait). Le regroupement en mots ou ensemble de mots se fait aussi via la couleur, les niveaux de gris, la largeur du stroke, des calculs géométriques comme l'espacement inter-caractères.

3.4.1. Méthodes basées sur la texture

Dans les images naturelles, le texte ait souvent une texture différente de son arrière-trame pour qu'il soit lisible. Cette caractéristique permet d'exploiter les méthodes d'analyse de texture pour l'extraction du texte tel que : Gaussian filter, wavelet decomposition, Fourier transform, local binary pattern (LBP) et Discret Cosin Transform (DCT).

Ces méthodes analysent l'image à plusieurs niveaux et extraient des caractéristiques qui seront par la suite classifiées, ce qui permettra la séparation des régions de texte de ceux ne comportant pas du texte.

L'utilisation de la texture permet la détection et la localisation même dans les images bruitées, mais le grand temps de calcul et la sensibilité à l'alignement du texte restent des inconvénients majeurs de cette classe.

3.4.2. Méthodes basées sur les régions

Les méthodes de cette classe utilisent des propriétés comme la couleur ou les niveaux de gris pour localiser le texte par rapport à l'arrière-trame. Cette classe peut être divisée à son tour en trois sous classes : les méthodes basées sur les composants connexes, celles basées sur les contours et celles basées sur les Strokes.

3.4.3. Méthodes basées sur les composants connexes

Généralement, ces méthodes regroupent de petits composants en d'autres plus grands au fur et à mesure jusqu'à l'émergement de toutes les régions de l'image. Ces dernières vont être classifiées en régions de texte/non texte suivant des calculs géométriques.

Les méthodes basées sur les composants connexes se composent de quatre étapes : le prétraitement, génération des composants connexes, filtrage et finalement le regroupement.

Ces méthodes ont l'avantage de simplicité d'implémentation, temps de calcul réduit, de possibilité d'utiliser leurs résultats directement à l'étape de reconnaissance de caractères, en plus qu'ils permettent la détection du texte de différentes tailles. Néanmoins, la vitesse est limitée par le grand nombre des régions qui doivent être analysées. Elles donnent un faible résultat si le texte ait la même couleur que l'arrière-trame. En plus, la position du texte doit être connue avant la segmentation.

3.4.4. Méthodes basées sur les contours

Les méthodes de cette classe utilisent les propriétés du texte qui doit être lisible, et par conséquent, va avoir de forts contours aux limites du texte avec l'arrière-trame. Ces méthodes utilisent un détecteur de contours suivi des opérations morphologiques pour la détection des régions de texte. La séparation des autres régions texte/non texte est faite via une classification ou des méthodes heuristiques.

L'utilisation des contours permet le développement des méthodes simples et efficaces dans le cas d'un texte ayant de forts contours et c'est ce qui limite leurs sucées. En effet, dans les conditions naturelles, c'est difficile d'avoir une bonne détection de contours.

3.4.5. Méthodes basées sur les Strokes

Le texte possède la propriété d'avoir une largeur de stroke similaire pour tous les caractères, ainsi les pixels de la même largeur peuvent être regroupés en composantes connexes. Les régions probablement contenant du texte sont extraites, vérifiées et enfin classifiées en composants de texte selon des règles géométriques.

Ces méthodes offrent la possibilité de détecter le texte à différentes tailles et orientations, mais la variation en luminance et l'arrière-trame complexe rend la détection des strokes (épaisseur de traits) une tâche très difficile.

3.4.6. Méthodes Hybrides

Les méthodes des deux classes décrites précédemment montrent des avantages et des limitations ainsi il sera utile d'exploiter l'avantage de l'une pour surmonter les limitations de l'autre.

Yao et al. [77] proposent un algorithme de détection de texte à orientation arbitraire en combinant la SWT utilisée généralement pour le texte aligné horizontalement avec des caractéristiques invariantes à la rotation. L'algorithme comporte quatre étapes : 1) L'extraction des composants : l'opérateur Canny est utilisé pour recouvrir les contours de l'image, le SWT est appliqué, et les composants connexes sont créés par des règles d'association simples. 2) Analyse des composants : Les éléments non-texte sont filtrés à l'aide d'un mécanisme de filtrage composé de deux couches, la première couche vérifie des propriétés statistiques et géométriques et filtre celles qui ne vérifient pas ces propriétés. Dans la deuxième couche, un classificateur filtre les régions qui ne comportent pas du texte selon des critères de différences des caractéristiques géométriques et texturales ; 3) Le chainage : les composants connexes ayant une hauteur et une largeur similaires sont ensuite regroupés de façon récursive. 4) Analyse des chaînes : les chaînes avec des probabilités inférieures à un certain seuil sont éliminées.

Chen et al. [78] utilisent Maximally Stable Extremal Regions (MSER) combiné avec l'opérateur Canny pour détecter les composants connexes pour lesquels la SWT est calculée et les lettres sont groupées en mots, néanmoins les arrières-trames complexes restent encore un défi. Koo et Kim [79] utilisent aussi MSER pour l'extraction des composants connexes, ensuite deux classificateurs sont utilisés. Un premier classificateur Adaboost est utilisé pour examiner les relations d'adjacences pour le regroupement des composants connexes. Un deuxième classificateur est utilisé pour filtrer les composants non texte. La méthode peut ne pas marcher pour le texte suffisamment incliné.

Pan et Hou [80] utilisent un détecteur de régions de texte pour localiser et déterminer l'échelle. Les C.C. candidats sont extraits via un algorithme de binarisation locale exploitant l'histogramme des gradients orientés des régions de l'image. Les régions non-texte C.C. sont filtrées en utilisant le modèle Conditional Random Field (CRF). Les composants non filtrés sont regroupés en blocs de texte avec une méthode de minimisation de l'énergie. Cette méthode donne de bons résultats pour les images de haute résolution qui n'est pas le cas des vidéos.

Epshtein et al. [12] présente un opérateur local appeler Stroke Width Transform (SWT) qui utilise la largeur constante du stroke pour extraire les régions contenant probablement du texte. Le stroke est une caractéristique des caractères. Il s'agit d'une partie de l'image formant une bande d'une largeur constante. Pour chaque pixel de l'image, la largeur du stroke auquel

appartient le plus probable est calculée et conservée comme une valeur du SWT. Les pixels ayant la même valeur du SWT sont regroupés en mots selon des règles géométriques en composants plus larges dites des lettres candidates. Cette méthode a l'avantage d'être indépendante du langage et de la taille du texte.

WANG et al. [81] combine la DCT, MSER pour la description et la détection du texte. Les DCT sont extraites et filtrées par taille. Deuxièmement, la classification se fait en combinant MSER avec descripteur de forme pour l'extraction des caractères individuels. Un processus de regroupement de caractères est appliqué en exploitant la séparation spatiale et la distance lexicale. Une autre étape de classification utilisant une fonction de distance est utilisée pour former les mots. Cette méthode repose sur le regroupement plutôt que l'apprentissage. Cependant, la vitesse est encore une limite quand on utilise plusieurs couches de regroupement. Rong et al. [82] présentent une méthode à deux niveaux pour l'extraction du texte. Le Stroke et le gradient sont utilisés pour détecter la polarité, qui sera utilisée pour l'extraction du texte.

3.5 Discussions

Les méthodes d'extraction d'information textuelle peuvent varier en principes, en stratégies et en caractéristiques utilisées. Cependant, jusqu'à présent, le problème d'extraction d'informations textuelles à partir des images naturelles reste non résolu vu l'absence d'une méthode globale permettant de répondre aux différentes variations de texte en taille, fonte, alignement et couleur, en plus des différentes contraintes imposées par l'environnement comme la variation de luminance et l'ombre. Ceci laisse ce domaine ouvert à de nouvelles recherches et contributions.

Les méthodes basées sur la texture présentent des résultats satisfaisants, mais nécessitent un parcours de l'image/trame à plusieurs niveaux, ce qui est très gourmand en temps d'exécution. En plus, la détection de texture est un problème délicat.

Les méthodes basées sur les régions offrent l'avantage de produire un résultat directement exploitable par les systèmes de reconnaissance de caractères avec moins de temps d'exécution que les méthodes basées sur la texture. Les composants connexes, les contours et les strokes sont des caractéristiques utilisées pour la détection des régions de textes.

Combiner deux méthodes ou plus permettra de bénéficier des avantages de l'une et de surmonter les limitations de l'autre. Cette combinaison semble une solution efficace, c'est le

cas des méthodes proposées récemment. En cas de besoin, des étapes de prétraitement ou d'amélioration sont nécessaires pour préparer les images résultantes d'une étape à autre et par conséquent augmenter les performances du système d'extraction d'informations textuelles.

Un autre point qui peut être exploité dans les documents vidéo, c'est la présence du texte dans un ensemble de trames successives, ce qui permet son suivi et l'analyse de son mouvement. Ceci parait utile pour les vidéos de faible qualité et d'améliorer les temps de réponse du système complet si le temps de suivi est moins que celui de la détection et la localisation, surtout lors de la manipulation des vidéos de grande taille.

3.6 Conclusion

Les informations textuelles incorporées dans les images sont une source importante d'informations permettant la description et la compréhension de leurs contenus sémantiques. Ces informations sont utilisées dans plusieurs domaines et applications.

Le processus d'extraction d'information textuelle est composé des cinq étapes : détection, localisation, segmentation, suivi et amélioration et enfin la reconnaissance des caractères. L'étape d'acquisition et de prétraitement à un impact important sur les résultats du système et doit être ajoutée au début des étapes du processus.

Plusieurs méthodes ont été proposées pour chaque étape, cependant le problème d'extraction d'informations textuelles reste non encore résolu, où l'absence d'une solution globale est dû aux variétés que présente le texte, en plus des conditions d'acquisition environnementales ce qui rend la tâche d'extraction plus délicate.

Partie II

Chapitre 04

Méthode récursive pour la segmentation et la structuration des vidéos de présentations

Récemment, les vidéos de présentations reçoivent plus d'attention. Elles sont de plus en plus utilisées pour le télé-enseignement. La quantité importante et croissante de vidéos disponibles nécessite des méthodes automatiques et efficaces pour l'indexation et la recherche à base de contenu. La structuration du contenu est une étape clé pour l'indexation. Parmi les niveaux structurels, le niveau segment est l'unité de structuration la plus utilisée pour l'indexation. De nature, les vidéos de présentations sont non structurées, et par conséquent, une étape d'analyse et de structuration est requise. De plus, les méthodes traditionnelles de segmentation vidéo ne peuvent pas être directement appliquées à ce genre de vidéos vu la nature de scènes homogènes et les différentes dégradations liées aux conditions d'acquisitions. Dans ce chapitre, nous introduirons notre première contribution. Nous présentons une méthode fiable et récursive de détection de points de transitions pour la structuration et la segmentation des vidéos de présentations. Pour une manipulation efficace, les images sont d'abord échantillonnées puis regroupées en segments. Par conséquent, au lieu de traiter chaque trame individuellement et de comparer chaque trame adjacente, nous traitons et comparons les segments adjacents. Pour l'extraction des caractéristiques et les mesures de similarité, la région de projection diapositive est initialement localisée, à partir de laquelle les caractéristiques sont extraites à l'aide des moments pseudo Zernike (MPZ). En fait, ces

moments sont invariants à l'échelle, à la rotation et à la translation, ce qui rend la méthode robuste contre les mouvements de caméra, la distorsion de perspective et les conditions de luminance. L'expressivité des MPZ est exploitée pour la mesure de similarité entre segments. En cas de dissemblance, le segment en cours est subdivisé récursivement en sous-segments jusqu'à ce que le point de transition exact soit localisé. L'approche montre qu'elle est robuste contre la luminance et la basse résolution. L'efficacité de la description des MPZ donne un taux d'erreur réduit.

4. Méthode récursive pour la segmentation et la structuration des vidéos de présentations

4.1 Introduction

L'indexation et la recherche automatiques de vidéos à base de contenu été un domaine de recherche actif pour la dernière décennie [23, 83]. Ces systèmes reposent sur l'organisation hiérarchique du contenu sous forme de segments séparés par des points de transitions pour permettre l'indexation et la navigation. Un segment est un ensemble de trames consécutives constituants une unité sémantique significative. Les segments sont considérés comme des unités structurelles fondamentales pour l'indexation vidéo et les plus adaptés pour la navigation et l'annotation à base de contenu [4].

Contrairement aux vidéos ordinaires, les vidéos de présentations sont généralement de contenu non structuré et ne permettent qu'une recherche séquentielle. Par conséquent, la vidéo doit être revue plusieurs fois en avant et en arrière pour rechercher une information particulière, ce qui est inefficace et long. Pour prendre en charge l'indexation et la recherche, leur contenu doit être segmenté en segments disjoints significatifs appelés *segments diapositives*. Un segment diapositive peut être vu comme une séquence de trames contenant la même diapositive. Un point de transition est détecté lorsqu'une transition de diapositive s'est produite. L'identification des transitions diapositives est une tâche très difficile vu que les vidéos: 1) sont généralement de mauvaise qualité en raison des conditions d'acquisition tel que le bruit, la variation de lumière, la distorsion et l'occlusion; 2) sont composées de scènes homogènes, où les changements dans la région de la diapositive ne montrent pas les changements significatifs de couleur; 3) contiennent le mouvement des conférenciers sur la région de projection, produisant un changement important pour la même diapositive.

La détection des points de transitions diapositives a été l'objet de nombreuses recherches [5, 49, 84]. Cependant, les solutions introduites fonctionnent dans des conditions spécifiques, ce qui ne couvre pas le cas de la plupart des vidéos et par conséquent l'absence d'une solution globale.

La détection des points de transitions (Shot Boundary Detection) ou la segmentation temporelle des vidéos se base sur la dissemblance entre les trames adjacentes. Le processus de détection de points de transitions comprend généralement trois étapes : l'extraction de caractéristiques, la mesure de similarité entre les trames adjacentes et enfin la détection. Dans la première étape, les caractéristiques visuelles sont extraites et des descripteurs sont générés pour les mesures de similarité. Parmi les caractéristiques visuelles de bas niveau, la forme est une caractéristique puissante incorporée dans presque tous les systèmes d'indexation et de recherche à base de contenu.

Les descripteurs de forme sont généralement divisés en deux catégories : les contours et les régions. Les approches basées sur le contour n'utilisent que les informations de contour tandis que les approches basées sur les régions prennent en compte les informations de contour et d'intérieur. Les fonctions de moments sont les descripteurs de forme les plus courants. En fait, l'efficacité des approches à base statistique, comme les moments, justifie leur large utilisation dans les applications de reconnaissance de formes et d'analyse d'images [1, 34, 85-87]. Selon l'étude de Tan et al. dans [86] les moments orthogonaux, tels que les moments de Zernike (MZ) [34] et les Moments Pseudo Zernike (MPZ) [8], présentent plus d'avantages que les autres types des moments. Parmi les moments orthogonaux, les moments Pseudo Zernike sont les plus immunisés contre le bruit et plus expressifs avec une redondance d'information minimale. Les MPZ présentent d'autres avantages importants tels que la robustesse au bruit, l'invariance de la rotation et de l'échelle, l'expressivité et la description des caractéristiques à plusieurs niveaux. Ils sont même plus efficaces que les moments de Zernike. Les MPZ ont été utilisés dans de nombreuses applications telles que l'analyse d'image, la reconnaissance de formes, l'indexation et recherche à base de contenu, la reconnaissance faciale, le tatouage.

Dans notre première contribution, nous proposons une approche robuste et récursive pour l'analyse structurelle. Les vidéos de présentations sont structurées en segments diapositives via une détection des points de transitions basée sur des moments pseudo-Zernike, résultant en un ensemble d'images clés.

La méthode proposée comprend principalement quatre étapes : échantillonnage et partitionnement des trames, détection et segmentation des régions diapositives, extraction des caractéristiques et mesures de similarité, et enfin, détection récursive des points de transition. L'échantillonnage et le partitionnement des trames consistent à ne prendre en compte que les trames dans une étape prédéfinie, du fait que les trames consécutives dans les vidéos de présentations ont des caractéristiques visuelles très similaires, cela évitera les comparaisons inutiles et réduit le taux d'erreur. Nous regroupons les trames échantillonnées en segments de *k* trames. La similarité entre segments successifs est calculée et comparée à un seuil. Si une dissemblance est trouvée, le segment actuel est subdivisé récursivement et vérifié pour trouver le point de transition exacte. Pour une extraction de caractéristiques fiable, les MPZ sont utilisés pour caractériser la région diapositive pour les mesures de similarité.

Les contributions principales de cette méthode sont :

1 / Puisque la configuration non stationnaire est supposée, la récupération de région diapositive est une étape importante. Cette étape améliore les performances globales du système en réduisant le temps de calcul et en limitant les régions de comparaisons. La méthode proposée permet la détection de la région diapositive en utilisant des MPZ.

2 / Extraction de caractéristiques robuste et précises pour les mesures de similarité. Les MPZ sont invariants à l'échelle, à la rotation et à la distorsion, ce qui rend la méthode robuste contre les différents bruits, les mouvements de la caméra, la distorsion de la perspective et les conditions de luminance ;

3 / Méthode récursive de détection de points de transitions diapositives par subdivision récursive de segments en sous-segments plus petits. Par conséquent, la recherche de point de transition considérera des demi-segments à chaque fois et les comparaisons de similarité inutiles sont évitées. Deux seuils sont combinés : Le seuil global est utilisé pour la vérification entre les segments et le seuil adaptatif local pour la recherche dans le segment.

4.2 Méthode récursive de détection de points de transitions diapositives

La méthode proposée consiste principalement en quatre étapes : Echantillonnage et partitionnement des trames, détection et extraction de région diapositive, extraction des caractéristiques et mesures de similarité, et enfin, détection récursive des points de transition.

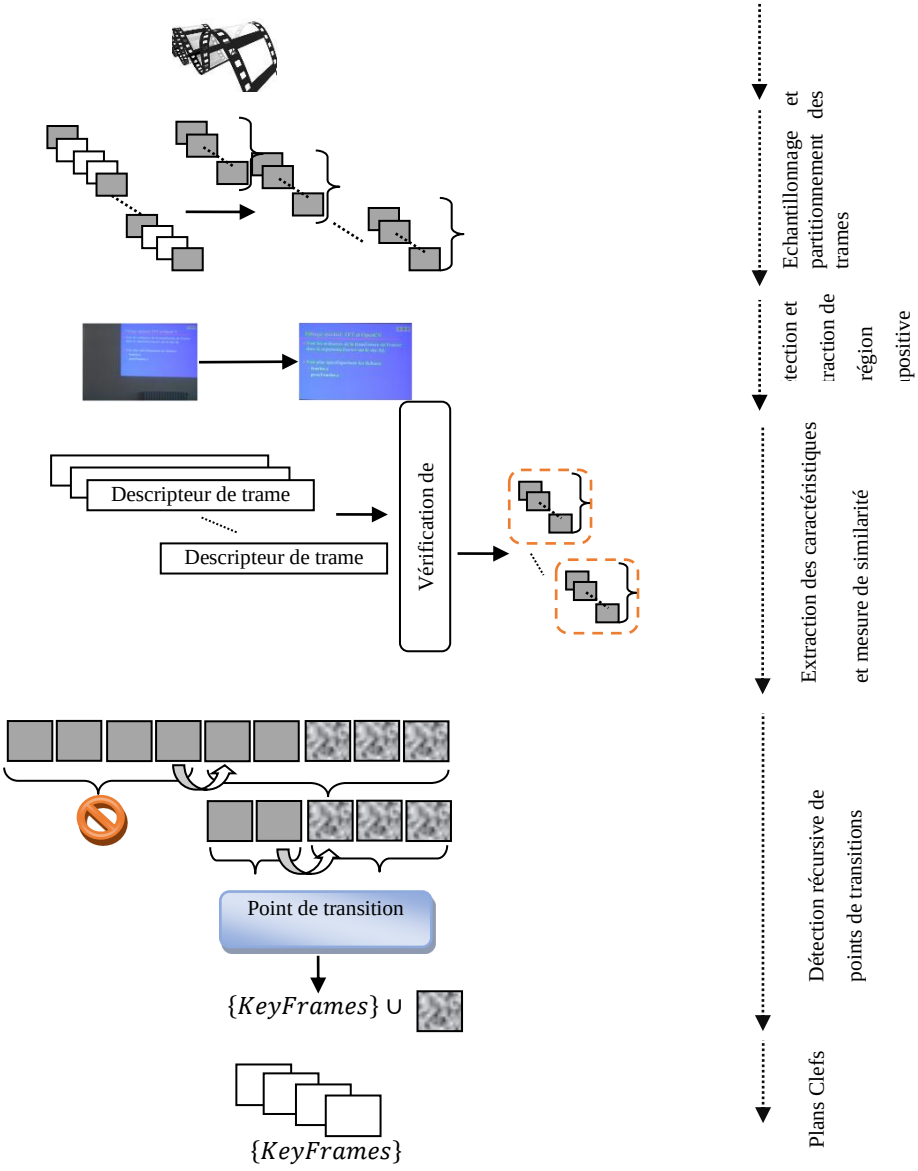

Fig. 4.1. Vue générale de la méthode proposée.

4.2.1. Echantillonnage et partitionnement des trames

Dans les vidéos de présentations, les trames consécutives ont des caractéristiques visuelles très similaires. Ainsi, les trames vidéo sont échantillonnés en ne prenant en compte que les trames à un pas prédéfini à partir de la première trame. Dans notre cas, le pas *Sp* est égal à 10. Par conséquent, pour une vidéo de *N* trames, l'ensemble F_{sample} considéré sera :

$$F_{sample} = \{ \cup_{i=1..N} F_i \,|i = i + Sp\} \quad (4.1)$$

où : F_i : est la trame numéro i.

Le partitionnement de la vidéo consiste à diviser F_{sample} en segments S_i consécutifs composés de K trames sans chevauchement. L'utilisation des segments de trames, au lieu de considérer chaque trame individuellement, est utile et améliore les performances globales du système en termes d'efficacité de calcul et de temps d'exécution. En effet, lors d'une présentation, l'intervenant aura besoin de temps pour dire quelque chose autour de la diapositive en cours ou interagir avec l'audience, aussi un peu de temps est perdu si la diapositive contient une animation. Ainsi quelques secondes seront nécessaires et aucune transition n'aura lieu dans ce délai. Par exemple, les vidéos de présentations peuvent être de 10 minutes ou plus, alors que la présentation ne peut contenir que 20 à 25 diapositives. Par conséquent, la méthode proposée utilise le segment vidéo S_i comme unité de traitement. L'intervalle de temps doit être suffisamment important pour éviter l'extraction et les mesures de similarité inutiles et suffisamment petit pour contenir une transition de diapositive. Dans notre cas k=140 trames consécutives, au cours des expérimentations, cette valeur est observée pour optimiser au mieux les résultats en respectant le compromis efficacité et temps de calcul.

4.2.2. Détection et extraction de la région diapositive

Dans les vidéos de présentations, la plupart des informations se localisent dans la région diapositive. Par conséquent, pour une description efficace et précise, nous ne considérons que la région diapositive. D'abord, nous avons détecté et extrait la région diapositive. Les autres objets de la scène tels que l'intervenant et l'audience sont considérés comme arrière-trame. Cependant, la région diapositive peut être occlue ou présente certaines distorsions dues aux mouvements. Les conditions d'acquisitions peuvent également provoquer un changement de luminance irrégulier affectant la région diapositive ce qui complique la tâche d'extraction. La détection et la segmentation de région diapositive se fait en deux étapes : d'abord détection via une segmentation à base de MPZ, suivi d'une étape de reconnaissance de forme pour l'extraction.

4.2.2.1. Détection de la région diapositive

La détection est appliquée via une segmentation à base des MPZ, qui consiste en deux étapes : (1) extraction des caractéristiques et (2) classification. La segmentation permet d'étiqueter chaque pixel par son appartenance à une région diapositive / non diapositive. Pour l'extraction des caractéristiques, nous utilisons les MPZ en tant que caractéristique locale à

partir d'une image HSV. Ensuite, l'étape de classification est effectuée en utilisant l'algorithme K-moyens pour k = 2. Le résultat est une image binaire où les pixels sont étiquetés par leur appartenance à la région diapositive / non diapositive (Fig. 4.2).

(a) (b) (c) (d)

Fig. 4.2. Détection de région diapositive. (a) trame originale, (b) Résultats de segmentation MPZ (c) Reconnaissance de forme (d) diapositive extraite.

4.2.2.2. Extraction de la région diapositive

Une étape de reconnaissance de forme est effectuée pour identifier la région diapositive. Généralement, la région diapositive à une forme carrée ou rectangulaire : cependant, l'angle de caméra peut provoquer une distorsion géométrique. La diapositive peut également être partiellement occlue. De manière générale, cette région est considérée comme un polygone quadruplé de forme régulière ; ainsi à partir de deux régions obtenues par classification, celle qui respecte cette propriété est considérée comme une région diapositive. La reconnaissance de forme est effectuée en détectant tout d'abord les contours à l'aide du détecteur de Canny, puis chaque contour est approximé et le plus grand quadruplet englobant la région est pris en tant que région diapositive. Enfin, pour des améliorations de précision, la trame originale est recadrée en fonction de la région reconnue et une nouvelle image est obtenue.

4.2.3. Extraction des caractéristiques et mesures de similarité

4.2.3.1. Extraction des caractéristiques

Une fois la région diapositive est segmentée et recadrée, elle est transformée en image de niveau de gris et les moments pseudo Zernike sont calculés en utilisant le même processus décrit ci-dessus. Les MPZ permettent une description efficace, robuste au bruit et invariante au changement d'échelle et à la rotation. L'expressivité, l'extraction multi-niveaux combinée avec l'utilisation d'une fenêtre glissante permet de capturer chaque petit changement dans la région diapositive. Pour une description efficace, nous utilisons un ordre élevé pour le calcul

des moments P_{max}=12. En effet, les moments orthogonaux permettent une description d'image par un ensemble de moments indépendants, chacun portant des informations différentes. Les moments de faible ordre décrivent les détails globaux de la région quant à ceux d'ordre élevé donnent plus d'informations locales. Le vecteur de caractéristiques global est obtenu par :

$$V = \left\{ \cup D_{x,y} \,\middle|\, x = 0 \ldots . \grave{N}, y = 0 \ldots . \grave{M} \right\} \quad (4.2)$$

où :

$\grave{N}$ et $\grave{M}$ sont le nombre de fenêtres en longueur et en largeur.

$D_{x,y} = \left\{ \cup_{P_{max}} f(x,y) \right\}$ est le vecteur descripteur de chaque fenêtre.

4.2.3.2. Mesures de similarité

Une fois que les vecteurs de caractéristiques sont calculés, la similarité est mesurée pour l'identification similaire / non similaire. Le segment identifié, pour qu'il contient probablement une transition, subie une autre étape récursive sinon la méthode passe au segment suivant.

La similarité entre deux trames est calculée en utilisant la distance euclidienne entre les vecteurs caractéristiques, puis le résultat est comparé au seuil τ calculé à base de la moyenne et les déviations des similarités calculées entre les trames du segment en cours. Le seuil τ est estimé d'une manière adaptative pour chaque segment. Selon [88] la moyenne μ et les écarts-types σ de tous les descripteurs de trames entre deux trames spécifiés peuvent être utilisés comme un seuil local et prend en compte les divers changements contextuels. Néanmoins, ceci est couteux et augmente la complexité et le temps de calcul. La méthode proposée implique un nombre réduit de trames dans le calcul de τ. Notez que deux seuils sont combinés ($G\tau$ et τ). Le seuil global $G\tau$ est utilisé pour la vérification entre les segments, tandis que le seuil adaptatif local τ est utilisé pour la recherche dans le segment. En effet, le seuil adaptatif introduit plus d'informations contextuelles, mais son estimation pour chaque segment est gourmande en temps de calcul et parfois inutile, par conséquent un seuil local n'est calculé que pour les segments candidats.

Néanmoins, le calcul de toutes les similarités entre toutes les trames d'un segment est couteux, ainsi, le calcul portera sur un nombre de trames égale à 12 choisis aléatoirement parmi les trames du segment en cours.

$$\tau = \mu(1 - 1/\sigma) \quad (4.3)$$

tel que : μ est la moyenne des distances entres les 12 trames du segment en cours.

σ est la déviation standard des 12 trames du segment en cours.

Les trames sont considérées comme dissemblables si la similarité est inférieure au seuil estimé. L'*Algorithme 4.1* décrit le processus suivi.

Algorithme 4.1. Mesure de similarité

Input: F_j^i , F_p^q numéros des trames **j, p** dans les fragments numéros **i, q**

Output: bool

Calculer μ pour le segment en cours ;

Calculer σ pour le segment en cours ;

$\tau = \mu(1 - 1/\sigma)$; /* seuil*/

Calculer la distance Euclidienne d *(frame1desciptor, frame2descriptor)*;

if $d < \tau$ then

Are_similar = true;

else

Are_similar = false;

End

4.2.3.1. Vérification de segment

Pour la mesure de similarité entre segments, il existe trois possibilités qui peuvent être reportés pour chaque segment :

a) Transition frontière : transition à la limite du segment actuel et au début du segment suivant.
b) Segment candidat : probablement contenant une transition.
c) Pas de transition.

Algorithme 4.2. Vérification de segment

Input: $\{Segment\ S_i\}$,$\{Keyframes\}$, ***k*** numéro d'une trame dans le segment.

Output: $\{Keyframes\}$.

```
for each segment Si do
    if Are_similar (F i k−1 , F i+1 0) then
        i=i+1;      /*pas de transition, aller au suivant*/
    else
     if Are_similar ( F i 0, F i k−1 ) then
            {Keyframes}= {Keyframes} ∪ F i+1 0 ;    /*transition*/
        else
            borninf =F i 0 ;
            bornsup = F i k−1 ;
            RecursiveTD (borninf, bornsup);
          end
      end
  end
```

La transition frontière est détectée si la similarité entre le dernier et la première trame de deux segments adjacents S_i et S_{i+1} est inférieure au seuil calculé Gτ. Si les trames sont dissemblables, la similarité entre la première et la dernière trames du segment en cours est calculée. S'ils sont similaires, la méthode passe au segment suivant, sinon S_i est marqué comme probablement contenant une transition et on passe à la méthode de détection récursive de point de transition. L'estimation du seuil global Gτ se fait par la moyenne des distances entre descripteurs des premières trames de tous les segments, ceci permet d'impliquer plus d'informations qui couvrent le long de la vidéo et par conséquent reflète au mieux son contenu.

La fonction RecursiveTD() utilisée dans l'Algorithme 4.2 représente la méthode de détection récursive des points de transition et sera définie dans l'Algorithme 4.3.

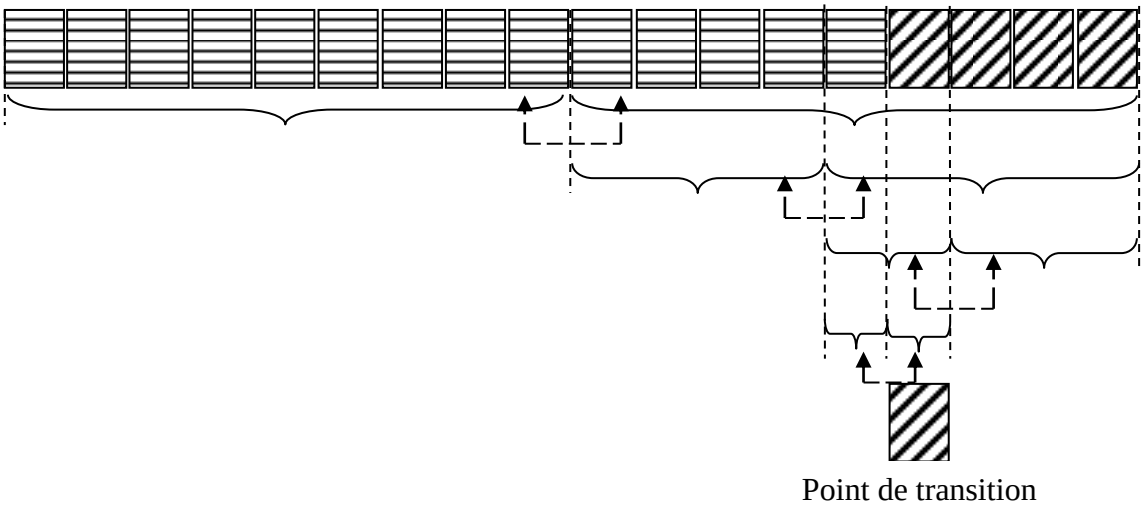

Fig. 4.3. Exemple d'exécution de la méthode de détection récursive des points de transitions.

4.2.4. Détection récursive des points de transitions

Chaque fois, un segment est marqué comme probablement contenant une transition. Une méthode de détection récursive de points de transition *RecursiveTD* est appliquée pour rechercher et détecter le point de transition exact et identifier les limites du segment. La trame qui représente un point de transition est considérée comme image clé et sera utilisée par notre deuxième contribution. Si aucune transition n'est trouvée, alors une fausse alerte sera signalée et le segment est ignoré. La méthode opère également comme une étape de vérification des résultats de l'étape précédente. L'ensemble des trames, constituants un segment, est divisé récursivement en deux intervalles plus petits jusqu'à trouver des trames adjacentes non similaires où la transition s'est produite, et le dernier est considéré comme une image clé. La recherche récursive évite l'extraction inutile des caractéristiques et les mesures de similarité entre toutes les trames d'un même segment. Par conséquent, elle permet une recherche efficace et rapide. L'Algorithme 4.3 décrit la méthode utilisée pour chaque segment. La figure (Fig. 4.3) montre un exemple d'exécution de la méthode proposée.

Algorithme 4.3. Détection récursive de point de transition (RecursiveTD)
Input: *borninf, bornsup ,* $\{Keyframes\}$.
Output: $\{Keyframes\}$.

```
if NOT Are_similar (bornsup, bornsup+1 ) then
   {Keyframes} = {Keyframes}∪ bornsup+1.
else
    if Are_similar (borninf, bornsup ) then
       Ignorer;
    else
       If  bornsup = borninf+1 then      /*deux trames récursives*/
            {Keyframes} = {Keyframes} ∪ bornsup;
       else
            RecursiveTD (borninf, borninf + (bornsup - borninf) div 2);
            RecursiveTD ((borninf + (bornsup - borninf) div 2) + 1, bornsup);
       end;
    end
end
```

La méthode récursive considère le segment marqué S_i composé de K trames comme intervalle [borninf, bornsup] où borninf = 0 et bornsup = k-1. L'intervalle est d'abord divisé en deux sous-intervalles [borninf, borninf + (bornsup - borninf) div 2] et [(borninf + (bornsup - borninf) div 2) +1, bornsup] puis la similarité est de comparer entre la première et la dernière trames de chaque segment. Si les trames sont similaires, le segment est ignoré, sinon il est subdivisé en deux sous-intervalles [borninf, borninf + (bornsup - borninf) div 2] et [(borninf + (bornsup - borninf) div 2) +1, bornsup] et le même processus est répété jusqu'à trouver l'intervalle composé de deux trames adjacentes dissemblables. Si, pour une sous-division, deux sous-intervalles sont ignorés, les trames sont considérées comme similaires et une fausse alerte est signalée, la méthode passe au segment étiqueté suivant.

4.3 Résultats et discussions

La méthode proposée est implémentée en C++ à l'aide de la bibliothèque OpenCV et les expériences sont effectuées sur un CPU à 2,50 GHz avec 2 Go de RAM, sous Windows 8.

4.3.1. Description de la base utilisée

Afin d'évaluer la méthode proposée, une base constituée de 12 vidéos de présentations a été utilisée. Les différentes vidéos reviennent à différents intervenants et sont de durée de 20 à 30 minutes et comprenant de 25 à 30 diapositives. Chaque vidéo a un modèle distinct et

comporte des variétés de contenus de diapositives, de faible résolution, avec une luminance non uniforme et une distorsion de perspective avec une configuration non stationnaire.

Le contenu de la base est classé en deux catégories selon le contenu des trames : vidéos sans intervenant, vidéos avec intervenant. Pour la première catégorie, de vidéos sans intervenant (C1), les trames ne contiennent que la région diapositive et un arrière-trame. Dans la deuxième catégorie, les vidéos avec intervenant (C2), chaque trame contient une région de diapositive, un arrière-trame et un intervenant, ce qui entraîne une occlusion de la région de diapositive. Six vidéos de chaque catégorie sont utilisées. La figure (Fig. 4.4**)** montre quelques exemples de trames des deux catégories.

(a)

(b)

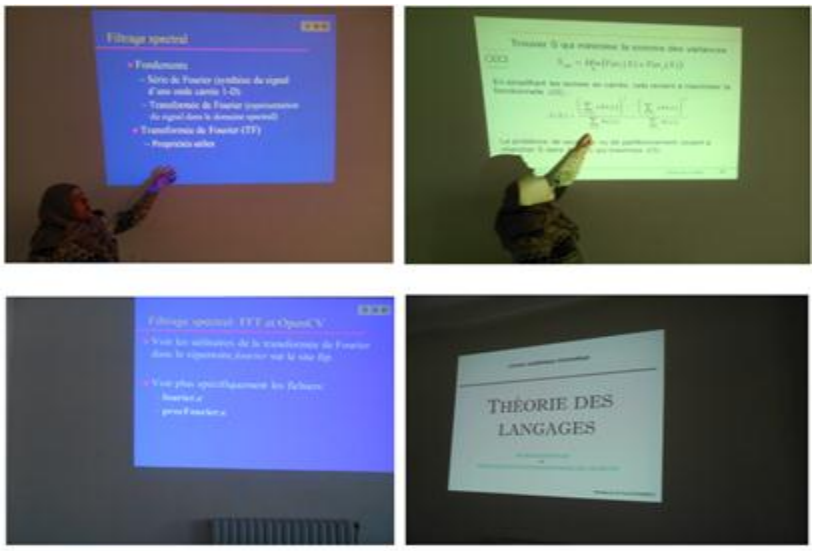

Fig. 4.4. Exemples de trames des deux catégories des vidéos utilisées. *(a)* catégorie C1, *(b)* catégorie C2.

4.3.2. Description des mesures utilisées

Pour l'évaluation de la méthode de détection de transition, plusieurs métriques ont étés utilisées : rappel, précision et score F [89]. La précision *P* est le rapport entre le nombre de transitions correctement détectées et le nombre de transitions détectées. Le rappel *R* est le rapport entre les transitions de diapositives correctement détectées et le nombre total de transitions de diapositives existantes. La F1 mesure combine *R* et *P* avec un poids égal de 0,5 en une seule mesure.

$$P = \frac{nombre\ de\ transitions\ correctement\ détectées}{Nombre\ total\ de\ transitions} \quad (4.4)$$

$$R = \frac{nombre\ de\ transitions\ correctement\ détectées}{nombre\ total\ de\ transitions\ détéctées} \quad (4.5)$$

$$F_1 = 2 * \frac{R*P}{R+P} \quad (4.6)$$

4.3.3. Evaluation

Les mesures sont appliquées sur chaque vidéo. Les résultats obtenus pour chaque vidéo sont donnés dans le Tableau 4.1.

Tableau 4.1. Résultats d'évaluation de la méthode proposée

Categories	Nombre total de transitions	Nombre de transitions correctement détectées	Nombre de transitions incorrectement détectées	Nombre de transitions incorrectement ignorées	Rappel	Précision	F-score
1	30	29	2	1	0,967	0,935	0,951
2	30	28	1	2	0,933	0,966	0,949
3	28	23	3	5	0,821	0,885	0,852
4	21	20	1	1	0,952	0,952	0,952
5	25	23	1	2	0,920	0,958	0,939
6	27	26	2	1	0,963	0,929	0,945
Total C1	**161**	**149**	**10**	**11**	**0,925**	**0,937**	**0,931**
7	23	20	4	3	0,870	0,833	0,851
8	22	21	2	1	0,955	0,913	0,933
9	25	19	4	6	0,760	0,826	0,792
10	25	20	3	5	0,800	0,870	0,833
11	30	26	1	4	0,867	0,963	0,912
12	21	19	1	2	0,905	0,950	0,927
Total C2	**146**	**125**	**15**	**21**	**0,856**	**0,893**	**0,874**
Total	**307**	**274**	**25**	**32**	**0,893**	**0,916**	**0,904**

Les résultats montrent que la méthode proposée a une précision élevée et un bon rappel pour les deux catégories. En C1, le score F1 est élevé. En C2, le taux de détection des erreurs est plus élevé que celui de C1. Les résultats sont principalement affectés par le mouvement de l'intervenant, où le locuteur occulte une grande partie de la région de la diapositive en plus des mouvements de la caméra tel que le zoom in et zoom out.

4.3.3.1. Comparaison avec d'autres méthodes existantes

Les résultats sont également comparés à d'autres méthodes, comme indiqué dans le Tableau 4.2**.** Pour la comparaison, des algorithmes récursifs [7] et [58] et autre non récursif [84] ont étés utilisés. Comme le montre le Tableau 4.2, la méthode proposée présente des performances élevées en termes de précision et rappel par rapport aux méthodes [7, 58, 84]. Les améliorations sont dues à l'expressivité, l'invariance des moments et le calcul récursif de la méthode proposées.

Tableau 4.2. Comparaison avec d'autres méthodes

	Méthode Proposée			**Méthode [7]**			**Méthode [58]**			**Méthode [84]**		
	R	**P**	**F**	**R**	**P3**	**F**	**R**	**P**	**F**	**R**	**P**	**F**
C1	0,925	0,937	0,931	0,912	0,901	0,906	0,896	0,868	0,882	0,878	0,799	0,837
C2	0,856	0,893	0,874	0,833	0,843	0,838	0,797	0,770	0,783	0,737	0,694	0,715
Total	0,893	0,916	0,904	0,873	0,872	0,872	0,847	0,819	0,832	0,808	0,747	0,773

Le Tableau 4.2 montre que :

- La méthode séquentielle [84] présente des résultats moins performants que les méthodes récursives [7, 11, 58]. La faible précision est causée par le changement de luminance qui influe sur les mesures de similarité et en résulte à un nombre important de fausses détections.
- La méthode récursive proposée donne des résultats meilleurs que celles des autres méthodes récursives [7, 58]. La variation de luminance et la composition de scène homogène est la cause principale des faibles résultats. Les deux méthodes reposent sur les valeurs des

pixels RGB pour mesurer les distances entre trames, ce qui mène à des fausses détections relatives aux changements de luminance et par conséquent une faible précision.

- Les méthodes récursives présentent un taux de rappel élevé. En effet, éviter les comparaisons inutiles réduit le nombre de détections et par conséquent réduit le taux d'erreur, mais d'autre part ceci influe aussi sur la précision où des transitions sont omises. Il est a noté que même pour le cas d'une recherche séquentielle, les performances sont fortement liées à la robustesse de description car une mauvaise description donne un nombre de détection élevé et un taux d'erreur élevé.
- Le cas des vidéos de C2, les quatre méthodes présentent des résultats moins intéressants que ceux de C1. Ceci est dû au nombre important de fausses détections relatives aux différents bruits : l'occlusion, zoom et au mouvement de la caméra.
- Les résultats que présente la méthode se justifient par le fait que la distorsion de perspectives et le mouvement de la caméra n'affecte pas les mesures de similarité. Ceci est dû à l'invariance à l'échelle et la rotation que présentent les moments PZ. Un autre avantage qu'on peut citer, la capacité de décrire les changements locaux et par conséquent permet de minimiser les fausses suppressions.
- La méthode proposée échoue s'il y a un grand changement dans la vue, ainsi de fausses détections sont reportées pour la même diapositive. Autre raison de fausses insertions est l'occlusion de la région diapositive par l'intervenant. Ce qui justifie pourquoi les résultats de la catégorie C2 sont moins que celles de C1.
- En conclusion, les résultats obtenus par la méthode proposée se justifient par le compromis entre l'effectivité de calcul et le nombre réduit de comparaisons.

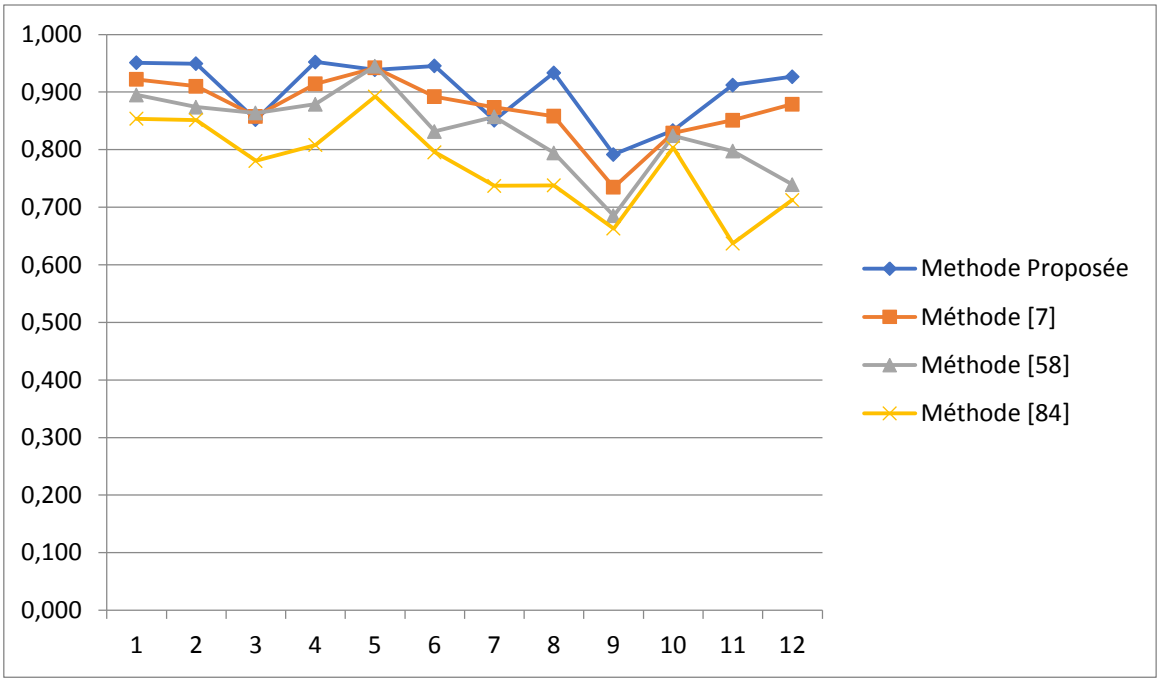

Fig. 4.5. Comparaison avec quelques méthodes existantes en F-mesure

La figure Fig. 4.5 montre que la méthode [7] donne de bons résultats qui sont assez similaires à ceux de la méthode proposée. Cependant, le temps d'exécution de la méthode proposée est significativement inférieur comme le montre la figure Fig. 4.6. Ceci est dû à plusieurs facteurs : le calcul rapide de PZM et la récursivité qui considère la moitié du segment à chaque tour. La méthode [7] nécessitent plus de temps de calcul car elle implique l'usage de multiples trames pour la génération du modèle, quant à la méthode [7], elle utilise les caractéristiques SIFT pour les mesures de similarités. En fait, même si le processus d'indexation est effectué hors ligne, le temps de calcul élevé pour une grande base de données reste une limitation majeure.

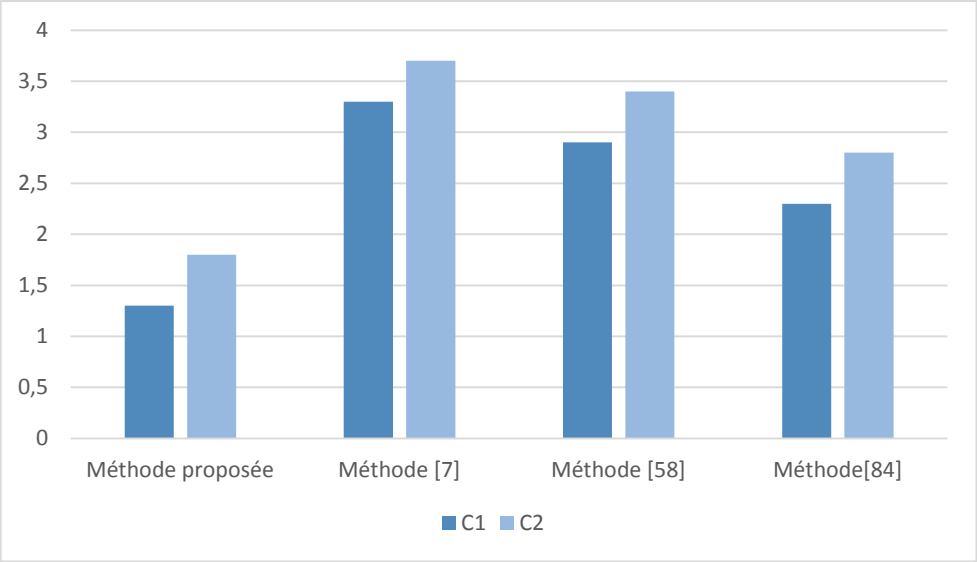

Fig. 4.6. Comparaison en temps de calcul avec les méthodes existantes

4.4 Conclusion

Ce chapitre présente une méthode de détection récursive de points de transitions pour la structuration et la segmentation des vidéos de présentations. Les moments pseudo Zernike sont utilisés comme caractéristiques pour la description et la mesure de similarité. Une méthode récursive de détection de points de transitions est appliquée si une dissemblance est trouvée entre segments. Le segment est divisé récursivement en sous-segments plus petits jusqu'à ce que deux trames différentes soient trouvées et qu'un point de transition soit détecté.

L'évaluation de la méthode proposée montre des performances élevées sur les deux catégories de vidéos que nous avons utilisé. Les expériences montrent que notre méthode possède des performances satisfaisantes en utilisant des vidéos de très basse résolution. Aussi, sa comparaison avec les méthodes existantes montre qu'elle présente des performances meilleures en termes de précision et en temps d'exécution. La subdivision récursive des segments a considérablement amélioré les performances en termes de complexité de calcul et de temps d'exécution. La faible complexité et le calcul rapide des moments favorise la méthode proposée vis à vis d'autres méthodes existantes.

Chapitre 05

Méthode de détection et localisation des informations textuelles

Les informations textuelles intégrées aux vidéos constituent un indice important pour l'indexation et la recherche à base de contenu. Néanmoins, ce texte de scène présente des caractéristiques difficiles à extraire. Ces difficultés sont principalement liées aux conditions d''acquisition et aux changements environnementaux résultants de la basse qualité des vidéos.

Dans ce chapitre nous introduisons notre deuxième contribution. Nous présentons une méthode de détection de texte de scène basée sur les moments Pseudo Zernike (MPZ) et les caractéristiques des stroke à partir des vidéos de présentations. La méthode consiste en trois étapes : Extraction de la région diapositive, segmentation du texte et filtrage.

Les résultats obtenus montrent que la méthode est robuste aux variations de luminance et à la basse résolution. L'efficacité des MPZ donne très peu de faux positifs en détection par rapport à d'autres approches existantes. De plus, les images résultantes peuvent être utilisées directement par les moteurs ROC sans aucun traitement supplémentaire.

5. Méthode de détection et localisation des informations textuelles

5.1 Introduction

De nos jours, la plupart des appareils mobiles [90] intègrent un appareil photo à haute résolution, ce qui permet une acquisition quotidienne, facile et importante d'images et de vidéos. La quantité croissante et l'utilisation fréquente des données multimédia émergent l'existence de technologies bien développées pour permettre l'accès rapide et efficace et le stockage à base de leur contenu. Le contenu visuel peut varier entre le texte, le mouvement ou le visage humain.

Le texte est un objet important capturé par la caméra. Il peut décrire le contenu et fournir des informations pour comprendre le contexte, il est utile pour l'indexation et l'annotation automatique. Des études [59] montrent que le texte est le premier objet détecté par la perception humaine et donne la plupart des informations contextuelles intuitives. En plus, il permet la compréhension de l'image et plus facile à extraire que d'autres contenus de haut niveau. Dans les environnements de télé-enseignement, le texte peut être utilisé dans diverses applications [49, 84, 91] tel que l'apprentissage en ligne et la vidéoconférence.

Les performances d'indexation sont liées au taux de Reconnaissance Optique de Caractères (ROC). Les moteurs de ROC sont développés pour opérer sur des images numérisées de haute résolution, incorporent du texte sur un arrière-plan uniform souvent blanc. Néanmoins, ce n'est pas le cas des images ou trames de scène naturelle où :

1/ Le texte est capturé dans l'image elle-même et fait partie des objets de la scène.

2/ Généralement les images sont de mauvaise qualité à cause du bruit produit par les conditions d'acquisition [92].

Par conséquent, ce genre d'images doit subir des étapes de traitement pour ressembler au plus aux images numérisées pour garantir un taux de reconnaissance élevé. Néanmoins, les images de scène présentent plusieurs conditions qui rendent le développement de telles méthodes de traitement une tâche délicate.

Le texte dans les images et les vidéos peut être une légende ou texte de scène. Le texte d'une légende est superficiellement ajouté aux images ou vidéos et superposé pour donner une description ou un résumé du contenu, tel que les nouvelles, les titres de films et les noms des

acteurs. Le texte de scène apparait naturellement dans les images et incorporé parmi les objets de la scène tels que les noms des joueurs et le texte dans les diapositives dans les vidéos de présentations. Les images numérisées de document à partir de documents imprimés peuvent également être considérées comme des images de texte de scène.

(a) (b) (c)

Fig. 5.1. Exemple du texte dans les images. (a) Document numérisé, (b) légende, (c) texte de scène. [59]

Le texte présente plusieurs propriétés : police, couleur, géométrique telles que : taille, alignement et espacement entre les caractères, mouvement, contours, distorsion de la perspective due à l'angle d'acquisition. Le texte de scène présente des propriétés supplémentaires en raison des conditions d'acquisition et les changements environnementaux telles que : faible résolution, luminance inégale, mouvement de la caméra et arrière-trame complexe qui rendent la détection et l'extraction une tâche extrêmement difficile.

Le système d'extraction d'informations textuelles (TIE) tel que défini dans [60] comprend cinq phases : détection et localisation, extraction, suivi, amélioration, binarisation et reconnaissance de caractères. La détection de texte est le besoin de connaître l'existence ou non de texte dans les images sans aucune information préalable sur la présence/absence du texte, ni sur l'emplacement, ni sur la taille ou même l'orientation. La localisation du texte est effectuée en localisant la région de texte détectée et en générant des cadres englobant autour des caractères ou d'un mot entier. L'extraction consiste à segmenter le texte localisé de l'arrière-trame et résulte en une image binaire. Généralement, une étape d'amélioration est nécessaire lorsque le texte est de mauvaise qualité. Enfin, un texte brut est généré à l'aide d'un moteur de reconnaissance optique de caractères (ROC).

Uchida dans [69] définit le TIE comme un processus à trois étapes : l'acquisition, la localisation et la reconnaissance. Qui n'est que le groupement des étapes décrites ci-dessus en

plus d'une étape supplémentaire : l'acquisition. En effet, les résultats de la détection, de la localisation et de la reconnaissance dépendent des conditions d'acquisition, ainsi, une étape supplémentaire doit être réalisée pour préparer les images/trames et améliorer les performances globales du système.

La détection de texte de scène a été considérée dans plusieurs recherches [59, 69, 79, 91] et différentes méthodes ont étés proposées. Cependant, jusqu'à présent, il n'il n'y a pas de solution globale. L'absence de telle méthode est due aux caractéristiques de texte de scène : bruit, faible résolution, luminance, grandes variétés d'alignement, polices et taille, distorsion de perspective, distorsion géométrique et arrière-trame complexe.

Dans notre deuxième contribution, nous présentons une méthode de détection et de segmentation de texte à partir des vidéos de présentation**.** En effet, le texte dans les vidéos de présentation est l'élément le plus important et le plus utilisé pour l'indexation à base de contenu, la navigation ou pour produire des supports de cours à partir des vidéos. Deux caractéristiques sont combinées pour une détection et une extraction efficace : les MPZ [8] et le stroke [12] **(**Fig. 5.2**)**. La segmentation à base des MPZ est utilisée pour la description des caractéristiques et la détection et localisation des régions texte. La caractéristique Stroke est extraite via l'opérateur local SWT [12] choisi pour son efficacité et sa capacité à extraire les composants connexes directement à partir des contours et utilisé pour le filtrages des régions non textuelles. La méthode se compose de trois étapes : la segmentation de la région diapositive, la détection et l'extraction des régions texte candidates et le filtrage.

Par rapport aux autres méthodes existantes, notre méthode :

1/ Permet à la fois la détection et la segmentation du texte directement à partir des trames de faible résolution sans aucune étape de prétraitement ;

2/ Les résultats peuvent également être utilisés directement pour la reconnaissance ;

3/ La détection et l'extraction de texte sont effectuées sans classificateur trainé. La propriété d'invariance de MPZ et l'utilisation d'informations multi-niveaux permettent une extraction de caractéristiques efficace ;

4/ Le calcul très rapide qui est très important pour les systèmes de recherche en temps réel. Et même pour les systèmes d'indexation des grandes archives des données.

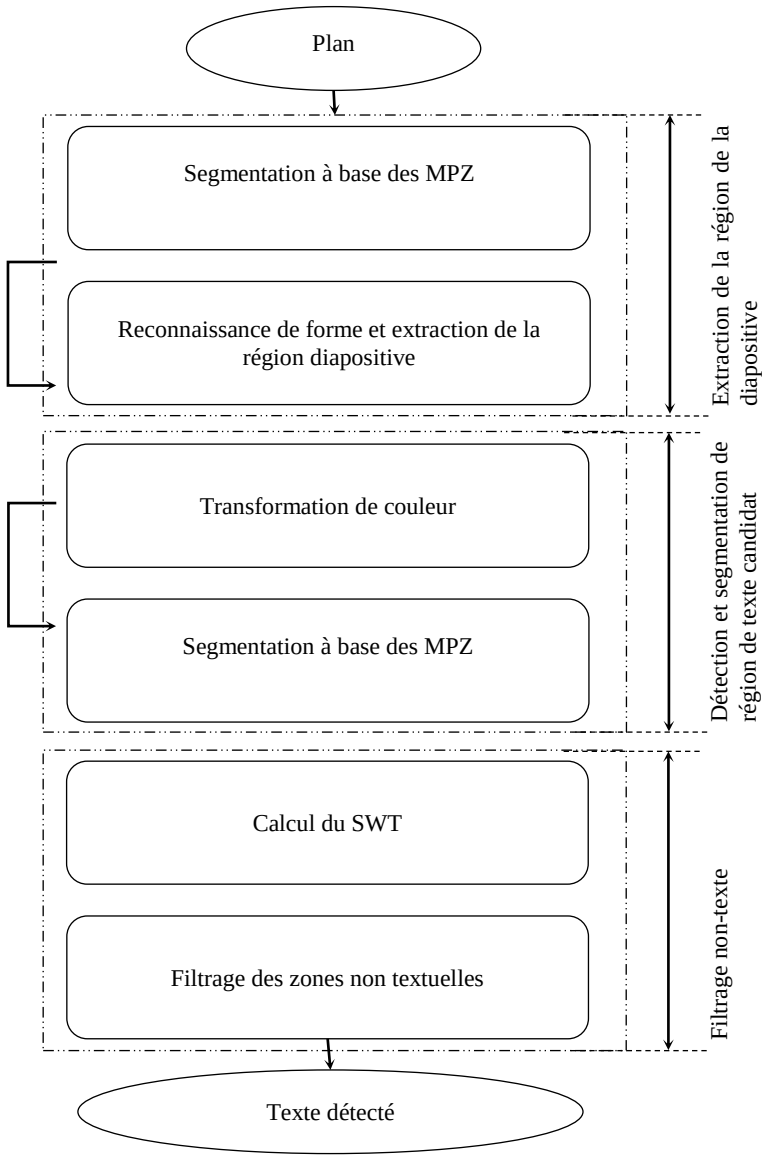

Fig. 5.2. Architecture de la méthode proposée.

5.2 Description de la méthode Proposée

5.2.1. Extraction de la région diapositive

L'objet principal dans l'environnement d'apprentissage est la région de projection diapositive où presque toutes les informations importantes se localisent. Le texte à l'intérieur est un composant clé pour l'indexation et la récupération vidéo. Par conséquent, la localisation de cette région est nécessaire pour l'amélioration des performances des systèmes d'extractions d'informations textuelles [93]. Pour les vidéos de présentations, la localisation et l'extraction de la région diapositive permet d'éliminer tous les autres objets de scènes non

utiles pour l'extraction du texte. Tout objet à l'extérieur de cette région sera considéré comme arrière-trame tels que l'intervenant et le public. Le reste du traitement sera effectué sur une partie plutôt que sur la trame entière.

L'extraction de la région diapositive est effectuée via une segmentation à base des MPZ sur des trames RGB. Trois étapes sont à effectuées : 1) partitionnement et normalisation de l'image, 2) extraction des caractéristiques, 3) regroupement, 4) extraction et recadrage des diapositives.

5.2.1.1. Partitionnement et normalisation

Les trois composantes R, G et B d'une trame sont extraites, chacune est divisée et sera partitionnée en fenêtres de taille égale W × W.

$$f^{x,y}(x_i, y_j) = f(Wx + x_i, Wy + y_i) \tag{5.1}$$

La fenêtre dans les images partitionnées peut être localisée par deux coordonnées *(x,y)* ou $x \in [0, NBlength - 1]$ et $y \in [0, NBwidth - 1]$. La fonction d'intensité d'image f au pixel(x_i, y_j) est donnée par l'équation :

La taille de la fenêtre W est estimée par des résultats expérimentaux, la valeur W = 6 donne un meilleur compromis entre temps d'exécution et qualité de la description.

Après le partitionnement de l'image, les coordonnées de chaque pixel sont normalisées à un espace de coordonnées polaires, où chaque bloc est mappé dans un cercle d'unité.

5.2.1.2. Extraction de caractéristiques

Les MPZ sont calculés pour toutes les fenêtres résultantes du descripteur de chacune. Le calcul des moments se fait via l'algorithme récursive proposé par [6]. Le choix est justifié par sa stabilité numérique et son temps de calcul rapide en calculant des termes factoriels récursifs [37, 93]. L'extraction de caractéristiques d'une image à base des MPZ est représentée sur la figure Fig. 5.3. Ce processus est appliqué pour chaque canal R, G et B.

5.2.1.3. Classification

Les descripteurs des fenêtres sont classés en utilisant l'algorithme k-moyens. Pour chaque fenêtre (i, j), les descripteurs obtenus à partir de chaque canal R, G et B sont combinés

pour former un descripteur global. Les blocs sont classés en fonction de leurs descripteurs en deux régions : région diapositive et arrière-trame.

5.2.1.4. Extraction

Une étape de reconnaissance de forme est effectuée pour identifier la région diapositive. En effet, la zone de projection diapositive est souvent d'une forme d'un polygone quadruplet (carrée ou rectangulaire). Le détecteur Canny est d'abord utilisé pour l'extraction des contours. Chaque contour est ensuite approximé et la plus grande forme de quadruplet est prise comme région diapositive.

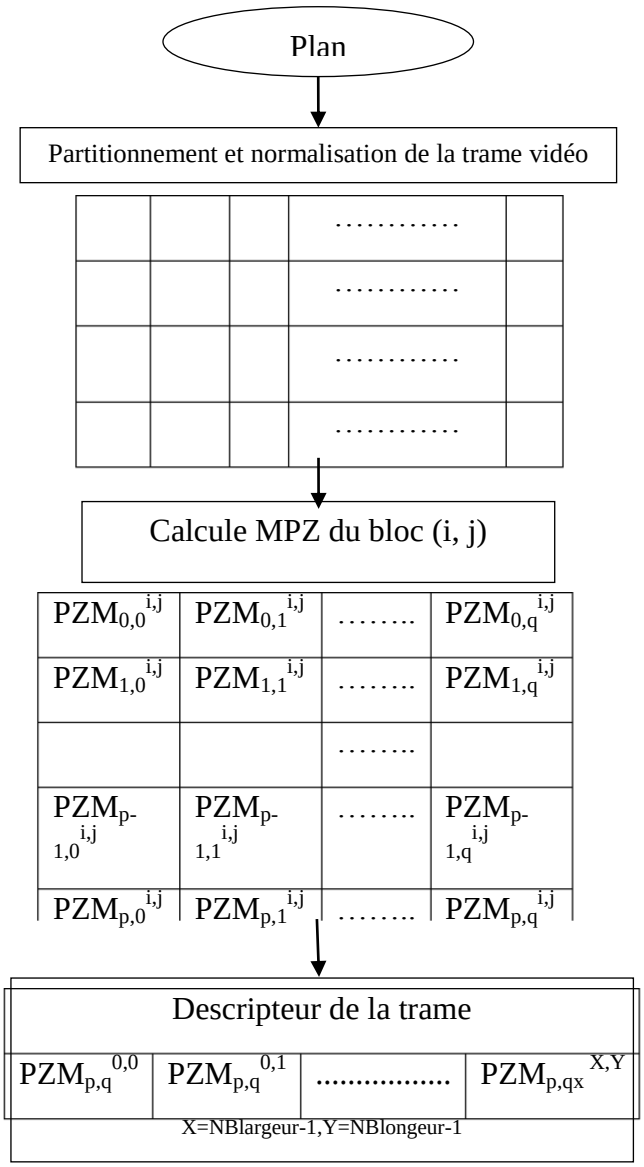

Fig. 5.3. Extraction des caractéristiques d'une image via les MPZ.

5.2.2. Segmentation de la région texte

La deuxième étape est la détection, la localisation et l'extraction des régions textes candidates en utilisant la segmentation à base des MPZ sur l'image recadrée.

L'image est d'abord convertie en espace colorimétrique HSV et le canal V est utilisé. Ensuite, une segmentation basée sur les moments Pseudo-Zernike est appliquée sur une image canal (Fig. 5.4). Le résultat est une image binaire dans laquelle les régions candidates sont segmentée à partir de l'arrière-trame.

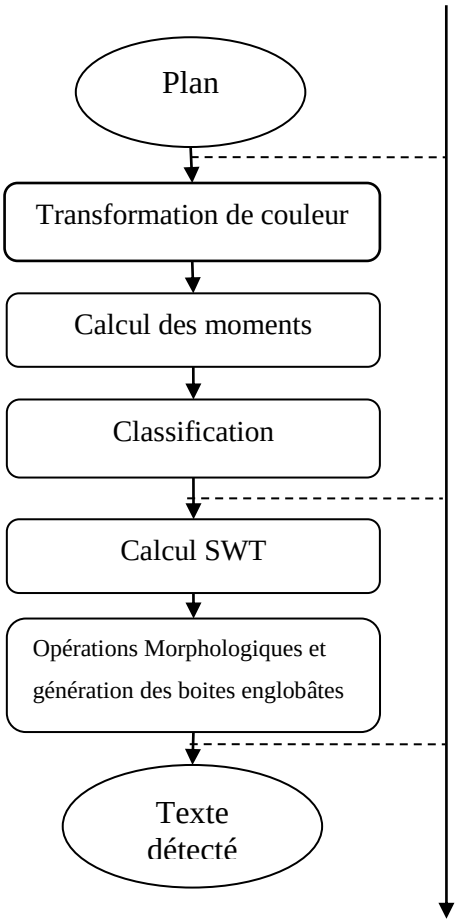

Fig. 5.4. Processus de segmentation des régions texte candidates et filtrage.

Cette étape est critique car le texte doit être bien récupéré et séparé de l'arrière-trame pour être utilisé comme entrée dans un moteur OCR pour la reconnaissance optique des caractères.

5.2.3. Filtrage

L'étape précédente permet la détection et l'extraction de la région de texte candidate. Pour filtrer les régions non-textuelles, la largeur de stroke est utilisée comme caractéristique. La largeur de stroke est récupérée via l'opérateur (SWT) [12]. La sortie SWT est un tableau de $N * M$ pixels où chaque pixel présente la largeur du stroke la plus probable auquel il appartient.

Les strokes sont filtrés en utilisant des propriétés géométriques : la hauteur et la largeur. Les régions de texte candidates sont englobées par des rectangles.

5.3 Résultats et discussions

La méthode proposée est implémentée en C++ à l'aide de la librairie OpenCV et les expériences sont réalisées sur un CPU à 2,50 GHz avec 2 Go de RAM sous Windows 8.

5.3.1. Description de la base utilisée

La méthode proposée a été testée sur une base de données contenant plus de 120 trames clés issues de l'étape de segmentation et structuration des vidéos de présentations. Les trames sont de basse résolution avec une variation de luminance élevée. Les trames contiennent des diapositives de différentes perspectives avec un arrière-trame non uniforme, coloré et texturé. Les régions diapositives contiennent à la fois du texte, des images et incorporent des symboles mathématiques. Le texte a une taille, une police, une forme, une orientation, une distorsion de perspective différentes. Quelques exemples de trames de la base de données sont présentés dans la figure (Fig. 5.5).

5.3.1. Mesures utilisées

Les mesures utilisées pour l'évaluation des performances sont le rappel *R* sur chaque trame, la précision *P* et la mesure *F-score*.

Pour l'évaluation, ces mesures sont appliquées sur chaque trame et les résultats sont présentés dans le Tableau 5.1. La région de texte est considérée comme correctement reconnue si le rapport d'intersection entre les boîtes englobantes calculées et celle du Ground Truth est plus de 80%.

Fig. 5.5. Exemples de trames de la base utilisée avec une faible résolution, une luminance non uniforme et une distorsion de perspective sur différents angles d'acquisition.

$$P = \frac{zones\ de\ texte\ correctement\ détectées}{zones\ de\ texte\ détectées} \tag{5.2}$$

$$\boldsymbol{R = \frac{zones\ de\ texte\ correctement\ détectées}{zones\ de\ texte\ totales}} \tag{5.3}$$

$$\boldsymbol{F = 2 \times \frac{R \times P}{R + P}} \tag{5.4}$$

5.3.2. Résultats et évaluation

Le Tableau 5.1 présente les résultats obtenus par la méthode proposée. Le rappel de 0.723 est le résultat de l'expressivité et la capacité de description des MPZ. D'autre part, la variation de luminance est éliminée via le résultat binaire de l'étape de détection de texte et par conséquent augmente le taux de détection des vrais positives et améliore les résultats de reconnaissance optique des caractères qui dépend de la qualité du texte en entrée.

Tableau 5.1. Evaluation de la méthode proposée

Précision	Rappel	F-mesure
0.861	0.723	0.785

Plan original	Extraction de la région diapositive	Détection et segmentation du texte	Filtrage

Fig. 5.6. Résultats de détection de texte. *(a)* Plans Originaux ; *(b)* Extraction de la région diapositive ; *(c)* Détection et segmentation des régions texte ; *(d)* Filtrage. Le texte détecté est encadré par des rectangles.

La précision élevée est justifiée par le fait d'utiliser le SWT pour détecter le texte de différentes tailles. En plus, la détection de la région diapositive limite la zone de recherche et par conséquent diminuer le taux d'erreur et optimiser le temps de calcul. La figure (Fig. 5.6) présente les résultats obtenus des différentes étapes de calcul de la méthode proposée.

La détection et la segmentation du texte sont validées à l'aide du moteur Tesseract OCR. Quelques résultats de reconnaissance optique des caractères sont donnés par la figure (Fig. 5.7**)**. Le résultat de cette étape peut servir pour la construction d'index permettant la recherche textuelle. Ce résultat permet aussi la construction d'une table de matière pour la navigation dans le document vidéo, en plus, ce résultat peut aussi être utilisé pour la préparation des supports de cours pour l'apprentissage distant.

Fig. 5.7. Exemples de résultats de reconnaissance. *(a)* Plan vidéo ; *(b)* Résultats de la reconnaissance optique des caractères.

5.3.3. Comparaison avec d'autres méthodes

Les résultats de la détection basée sur les MPZ [94] sont comparés aux autres descripteurs de régions. Pour comparaison, deux types de moments distincts, orthogonaux et non orthogonaux : Les moments de Zernike [34] et les moments de HU [39] sont utilisés pour l'extraction des caractéristiques.

La comparaison est faite en termes d'efficacité (rappel et précision) et de temps de calcul. Certains résultats sont donnés sur la figure Fig. 5.8. Le Tableau 5.2 montre qu'en

général, les moments orthogonaux donnent de meilleurs résultats que les moments non-orthogonaux. Les résultats donnés par les moments MPZ et Zernike sont très similaires. Cependant, la méthode à base des MPZ présentent un temps de calcul meilleur.

Plan original MPZ MZ HU

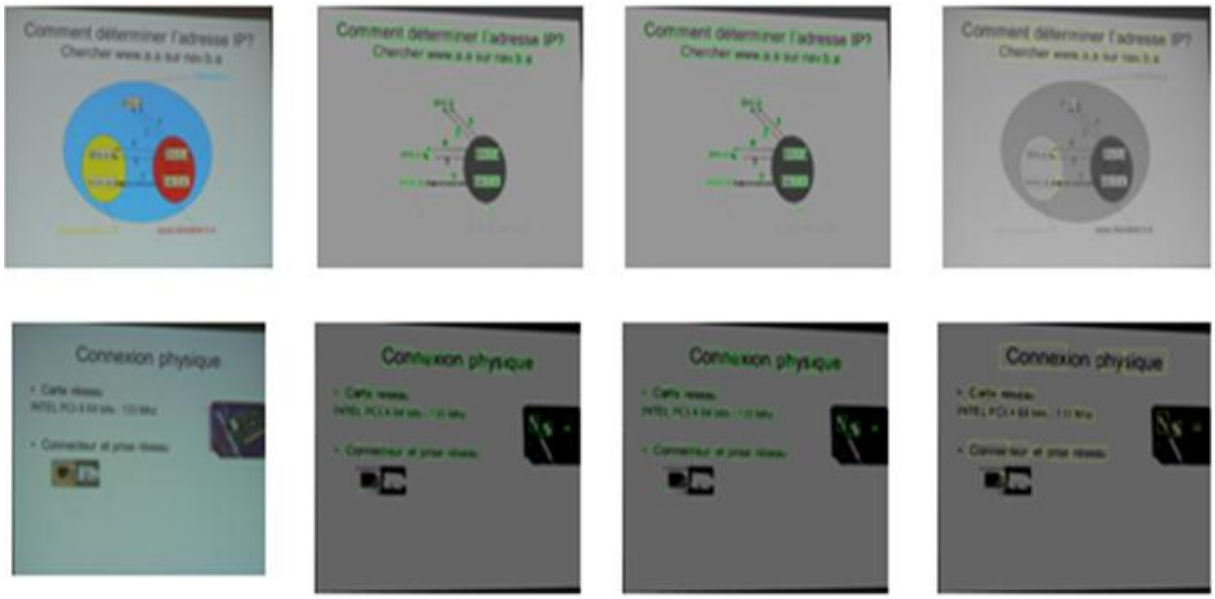

Fig. 5.8. Résultat de détection de texte via les différents moments.

Tableau 5.2. Comparaison des performances des différents moments.

	Précision	**Rappel**	**Temps d'exécution moyen**	
			Extraction de caractéristique	Détection de texte
PZM	0.861	0.723	0.10 ms	0.31 ms
Hu	0.709	0.680	0.28 ms	0.46 ms
ZM	0.834	0.720	0.33 ms	0.52 ms

Les résultats obtenus par la méthode proposée sont comparés avec ceux des méthodes existantes appliquées à des vidéos de présentations. Parmi les méthodes existantes, nous avons comparé avec deux d'entre elles: La méthode de Wang [84] et celle de Merler [91]. Les

résultats obtenus par les différentes méthodes sont comparés en termes d'efficacité (rappel/ précision). Le

Méthodes	Précision	Rappel	F-mesure
Méthode proposée	0.861	0.723	0.785
Wang [84]	0.709	0.680	0.694
Merler [91]	0.834	0.720	0.773

Tableau 5.3 et les figures Fig. 5.9 et Fig. 5.10 montrent les améliorations apportées par la méthode proposée par rapport aux autres méthodes. Les améliorations sont dues au taux d'erreur réduit, ceci est dû à 1) la détection robuste de la région diapositive qui diminue le taux des fausses détections et par conséquent une précision plus élevée et 2) l'invariance et la capacité de description multi-niveaux des moments PZ. La région diapositive et le texte ont été détectés et segmentés avec précision, et le nombre de fausses détections par la méthode proposée est faible par rapport aux autres méthodes.

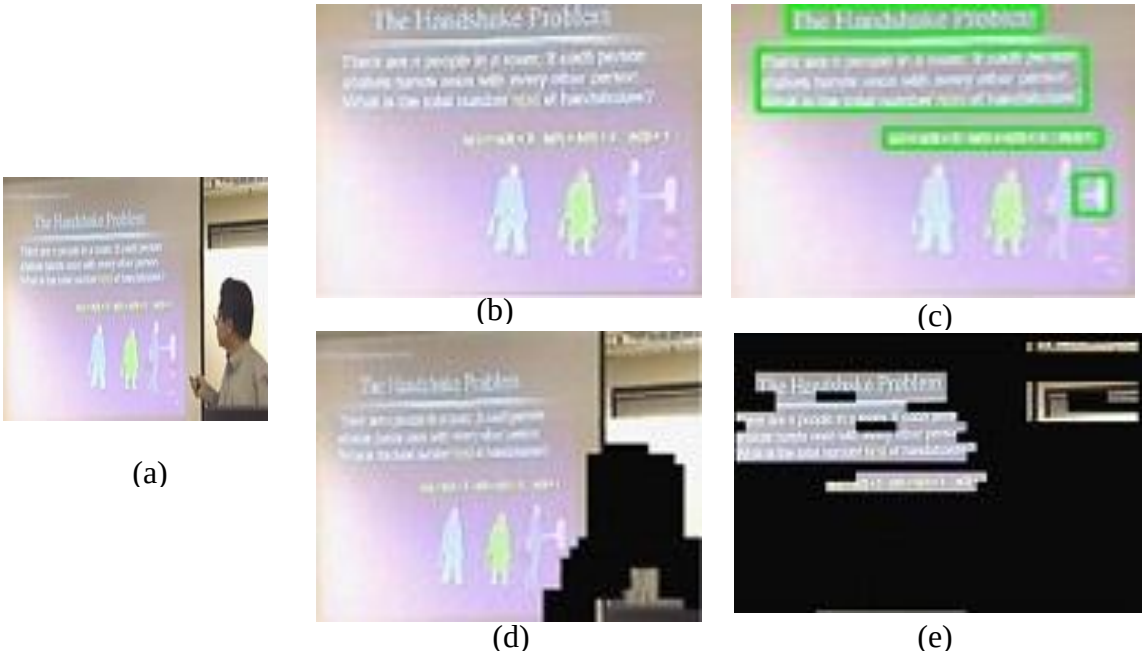

Fig. 5.9. Comparaison des résultats de détection de texte. *(a)* trame originale ; *(b)* résultat de l'extraction de la région diapositive par la méthode proposée; (c) résultat de détection du texte par la méthode proposée ; (d) résultat de la détection de la région diapositive via la méthode de Wang [84]; (e) résultat de détection de texte via la méthode de Wang [84].

Fig. 5.10. Comparaison des résultats de détection de texte. *(a)* trame originale ; *(b)* résultat de détection via la méthode de Merler [91] ; *(c)* résultat de détection via la méthode proposée.

(a) (b) (c)

Tableau 5.3. Comparaison des performances avec d'autres méthodes.

Méthodes	Précision	Rappel	F-mesure
Méthode proposée	0.861	0.723	0.785
Wang [84]	0.709	0.680	0.694
Merler [91]	0.834	0.720	0.773

5.4 Conclusion

Une méthode de détection et d'extraction des informations textuelles à partir de vidéos de présentations a été proposée. En premier temps, la méthode détecte et extrait la région de projection diapositive comme étape de prétraitement pour limiter la zone de traitement à la région d'intérêt. Ensuite, la région texte est détectée et segmentée en utilisant les moments Pseudo Zernike. Enfin, l'étape de filtrage à l'aide de SWT et les opérations morphologiques sont appliquées. Nombreux problèmes sont élevés : les conditions d'acquisition, la luminance, l'arrière-trame complexe et les caractéristiques textuelles.

Les expériences montrent de bonnes performances sur des trames de vidéos de présentations. Comparativement aux méthodes existantes, la méthode proposée montre de meilleurs résultats en termes d'efficacité. De plus, la méthode proposée permet à la fois la détection de texte et la segmentation, les résultats peuvent être utilisés directement pour la reconnaissance de caractères.

Conclusion Générale et perspectives

Dans ce travail nous avons abordé le problème de structuration et de caractérisation du contenu des vidéos de présentations. L'objectif était d'extraire des informations pertinentes permettant la description et la compréhension de leurs contenus sémantiques. L'indexation et la recherche des vidéos de présentations nécessitent de passer en premier temps par l'organisation structurelle de leurs contenus.

Nous avons présenté l'ensemble des techniques d'analyse structurelle et d'extraction des informations textuelles à partir des documents vidéo, ainsi que les méthodes dédiées aux vidéos de présentation. Ce type de vidéos souffre de plusieurs problèmes : L'homogénéité des scènes et le contenu long et non structuré, en plus de la basse qualité causée par les différentes dégradations relatives aux conditions d'acquisition. Comme présenté dans la première partie, les solutions actuelles souffrent d'êtres spécifiques à des conditions prédéfinies et par conséquent l'absence d'une solution globale.

Notre première contribution s'entoure autour de la segmentation et l'organisation du contenu visuel des vidéos de présentation. Une méthode robuste et récursive permettant d'identifier les segments diapositives a été proposée. L'identification est faite via la détection de points de transitions entre les différents segments via une fonction qui divise tout segment candidat en sous-segments jusqu'à l'arrivée au point exact ou la transition apparaît.

Dans la deuxième contribution, une méthode de détection, localisation et extraction des informations textuelles naturellement enfouies dans les trames de la vidéo est introduite. Ce texte fait partie des objets de la scène, il subit plusieurs dégradations relatives aux conditions

d'acquisition et aux changements environnementaux. Nous exploitons l'efficacité des moments orthogonaux pseudo Zernike pour l'extraction des régions texte candidates. Suivi d'une étape de filtrage appliquée via l'opérateur SWT et des opérations morphologiques.

Pour optimiser le traitement, dans les deux contributions, une étape de prétraitement a eu lieu. Il s'agit de l'étape d'identification et de segmentation de la région de projection diapositive. C'est la région d'intérêt qui contient la plupart des informations utiles. La région de projection est retenue pour les traitements futures quant aux autres objets de la scène seront ignorés.

Les méthodes proposées ont été testées sur deux catégories des vidéos : ceux comprenant les diapositives seulement et ceux contenant les diapositives avec un intervenant. Bien que nous travaillons sur des trames de faible résolution, des résultats satisfaisants ont était obtenus. Nos méthodes montrent des résultats plus performants par rapport à d'autres méthodes récentes. Ceci est dû à l'invariance des MPZ. Aussi, leurs capacités de description multi niveaux a mené à avoir un taux d'erreur minime. La faible complexité et le calcul rapide de ces moments est un avantage majeur pour tout système d'indexation, car même si l'indexation est faite en hors ligne, le temps d'extraction des caractéristiques, surtout pour des grandes archives de documents multimédias, est critique. La méthode récursive a amélioré considérablement le temps et la complexité du traitement.

Perspectives

Le travail présenté se concentre sur la segmentation et l'extraction d'informations visuelles à partir des vidéos de présentations. Il s'agit de la première étape du processus d'indexation à base de contenu. Comme perspectives nous prévoyons :

- L'exploitation des résultats pertinents obtenus pour l'étape d'indexation et de recherche à base de contenu.
- La méthode détecte les transitions en avant (forward transition). Nous prévoyons l'extension de la méthode pour la détection de transitions en arrière (backward transitions).
- Notre objectif était de structurer selon l'aspect temporelle un document vidéo à partir des informations textuelles enfuies dedans. Le contenu textuel extrait peut aussi servir pour la structuration à base de contenu des vidéos via la construction de tables des matières.

REFERENCES

1. Pedrotti, M. and N. Nistor. Online Lecture Videos in Higher Education: Acceptance and Motivation Effects on Students' System Use. in 2014 IEEE 14th International Conference on Advanced Learning Technologies. 2014.

*2. Furini, M., On introducing timed tag-clouds in video lectures indexing. Multimedia Tools and Applications, 2018. **77**(1): p. 967-984.*

3. Grünewald, F., et al. Next Generation Tele-Teaching: Latest Recording Technology, User Engagement and Automatic Metadata Retrieval. 2013. Berlin, Heidelberg: Springer Berlin Heidelberg.

*4. Hu, W., et al., A Survey on Visual Content-Based Video Indexing and Retrieval. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011. **41**(6): p. 797-819.*

*5. Yang, H. and C. Meinel, Content Based Lecture Video Retrieval Using Speech and Video Text Information. IEEE Transactions on Learning Technologies, 2014. **7**(2): p. 142-154.*

6. Sadiq, A.-R.M., Fast computation of pseudo Zernike moments. Vol. 5. 2010, Heidelberg, ALLEMAGNE: Springer. 8.

*7. Jeong, H.J., et al., Automatic detection of slide transitions in lecture videos. Multimedia Tools Appl., 2015. **74**(18): p. 7537-7554.*

*8. Teh, C.H. and R.T. Chin, On image analysis by the methods of moments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1988. **10**(4): p. 496-513.*

*9. Zhu, Y., C. Yao, and X. Bai, Scene text detection and recognition: recent advances and future trends. Frontiers of Computer Science, 2016. **10**(1): p. 19-36.*

*10. Yin, X.C., et al., Text Detection, Tracking and Recognition in Video: A Comprehensive Survey. IEEE Transactions on Image Processing, 2016. **25**(6): p. 2752-2773.*

11. Belkacem, S., L. Guezouli, and S. Zidat, Pseudo Zernike moments based approach for text detection and localization from lecture videos. International Journal of Computational Science and Engineering (IJCSE), 2016.

12. Epshtein, B., E. Ofek, and Y. Wexler, Detecting text in natural scenes with stroke width transform, in CVPR. 2010, IEEE. p. 2963-2970.

13. Clark, R.C. and R.E. Mayer, e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning. 2016: Wiley.

*14. Arkorful, V. and N. Abaidoo, The role of e-learning, advantages and disadvantages of its adoption in higher education. International Journal of Instructional Technology and Distance Learning, 2015. **12**(1): p. 29-42.*

15. Benraouane, S.A., Guide pratique du e-learning: Stratégie, pédagogie et conception avec le logiciel Moodle. 2011.

16. Mielnikoff, M. Qu'est-ce que l'E-Learning? 2005.

17. Simonson, M., S. Smaldino, and S.M. Zvacek, Teaching and Learning at a Distance: Foundations of Distance Education, 6th Edition. 2014: Information Age Publishing.

18. Tuna, T., et al. Topic based segmentation of classroom videos. in 2015 IEEE Frontiers in Education Conference (FIE). 2015.

19. Marques, O. and B. Furht, Content-Based Image and Video Retrieval. 2002: Springer US.

20. Furht, B. and O. Marques, Handbook of Video Databases: Design and Applications. 2003: CRC Press.

21. Yang, H., C. Oehlke, and C. Meinel. An Automated Analysis and Indexing Framework for Lecture Video Portal. 2012. Berlin, Heidelberg: Springer Berlin Heidelberg.

22. Yang, H., et al. Lecture Video Indexing and Analysis Using Video OCR Technology. in 2011 Seventh International Conference on Signal Image Technology & Internet-Based Systems. 2011.

*23. Balasubramanian, V., S.G. Doraisamy, and N.K. Kanakarajan, A multimodal approach for extracting content descriptive metadata from lecture videos. Journal of Intelligent Information Systems, 2015. **46**: p. 121-145.*

24. SenGupta, A., et al. Video shot boundary detection: A review. in 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). 2015.

25. Amato, F., et al., Content-Based Multimedia Retrieval, in Data Management in Pervasive Systems, F. Colace, et al., Editors. 2015, Springer International Publishing: Cham. p. 291-310.

26. Ma, X., et al. Content based Video Retrieval, Classification and Summarization: The State-of-the-Art and the Future. 2009.

*27. Li, K., et al., Structuring Lecture Videos by Automatic Projection Screen Localization and Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015. **37**: p. 1233-1246.*

28. Adcock, J., et al., TalkMiner: a lecture webcast search engine, in Proceedings of the 18th ACM international conference on Multimedia. 2010, ACM: Firenze, Italy. p. 241-250.

29. Sack, H. and J. Waitelonis. Exploratory Semantic Video Search with yovisto. in 2010 IEEE Fourth International Conference on Semantic Computing. 2010.

30. Gigonzac, G., F. Pitie, and A. Kokaram. Electronic slide matching and enhancement of a lecture video. in 4th European Conference on Visual Media Production. 2007.

*31. Medhi, S., C. Ahmed, and R. Gayan, A Study on Feature Extraction Techniques in Image Processing. International Journal of Computer Sciences and Engineering, 2016. **4**(7): p. 89-93.*

*32. Ansari, M.A. and M. Dixit, An Image Retrieval Framework: A Review. International Journal of Advanced Research in Computer Science, 2017. **8**(5): p. 692-699.*

33. Kumar, G. and P.K. Bhatia. A Detailed Review of Feature Extraction in Image Processing Systems. in 2014 Fourth International Conference on Advanced Computing & Communication Technologies. 2014.

*34. Teague, M.R., Image analysis via the general theory of moments. Journal of the Optical Society of America, 1980. **70**(8): p. 920-930.*

*35. Kanan, H.R., K. Faez, and Y. Gao, Face recognition using adaptively weighted patch PZM array from a single exemplar image per person. Pattern Recogn., 2008. **41**(12): p. 3799-3812.*

*36. Haddadnia, J., K. Faez, and M. Ahmadi, An Efficient Human Face Recognition System Using Pseudo Zernike Moment Invariant and Radial Basis Function Neural Network. IJPRAI, 2003. **17**: p. 41-62.*

37. Papakostas, G.A., et al., Efficient computation of Zernike and Pseudo-Zernike moments for pattern classification applications. Vol. 20. 2010, Heidelberg, ALLEMAGNE: Springer. 9.

38. Shah R, Zimmermann R (2017) Lecture Video Segmentation. In Multimodal Analysis of User-Generated Multimedia Content pp: 173-203.

*39. Hu, M.-K., Visual pattern recognition by moment invariants. IEEE Transactions on Information Theory, 1962. **8**(2): p. 179-187.*

*40. Atrevi, D.F., et al., A very simple framework for 3D human poses estimation using a single 2D image: Comparison of geometric moments descriptors. Pattern Recognition, 2017. **71**: p. 389-401.*

41. Wang, X.-Y. and L.-M. Hou, A new robust digital image watermarking based on Pseudo-Zernike moments. Vol. 21. 2010, Heidelberg, ALLEMAGNE: Springer. 18.

42. Dai, X., et al., Pseudo-Zernike Moment Invariants to Blur Degradation and Their Use in Image Recognition, in IScIDE, J. Yang, F. Fang, and C. Sun, Editors. 2012, Springer. p. 90-97.

*43. Hiremath, P.S. and J. Pujari, Content based image retrieval using color boosted salient points and shape features of an image. International Journal of Image Processing, 2008. **2**(1): p. 10-17.*

44. Ansari, A., et al. Content based Video Retrieval Systems - Methods, Techniques, Trends and Challenges. 2015.

*45. Yuan, J., et al., A Formal Study of Shot Boundary Detection. IEEE Transactions on Circuits and Systems for Video Technology, 2007. **17**(2): p. 168-186.*

46. Poleg, Y., C. Arora, and S. Peleg, Temporal Segmentation of Egocentric Videos, in Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014, IEEE Computer Society. p. 2537-2544.

47. Mandal, S.C.S.G.S.R.K. Emerging ICT for Bridging the Future. in the 49th Annual Convention of the Computer Society of India (CSI)

2014. India: Springer International Publishing.

48. Vora, C., B.K. Yadav, and S.S. Sengupta. Comprehensive Survey on Shot Boundary Detection Techniques. 2016.

49. Chong-Wah, N., P. Ting-Chuen, and T.S. Huang. Detection of slide transition for topic indexing. in Proceedings. IEEE International Conference on Multimedia and Expo. 2002.

50. Mukhopadhyay, S. and B. Smith, Passive capture and structuring of lectures, in Proceedings of the seventh ACM international conference on Multimedia (Part 1). 1999, ACM: Orlando, Florida, USA. p. 477-487.

*51. John, C., A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986. **8**(6): p. 679-698.*

52. Porter, S.V., Video Segmentation and Indexing using Motion Estimation. 2004, University of Bristol.

53. Zhang, H.J., A. Kankanhalli, and S.W. Smoliar. Automatic partitioning of video. in Multimedia systems. 1993.

*54. Smeaton, A.F., P. Over, and A.R. Doherty, Video shot boundary detection: Seven years of TRECVid activity. Comput. Vis. Image Underst., 2010. **114**(4): p. 411-418.*

55. Zhao, Z.-C. and A.-N. Cai. Shot Boundary Detection Algorithm in Compressed Domain Based on Adaboost and Fuzzy Theory. 2006. Berlin, Heidelberg: Springer Berlin Heidelberg.

*56. Liuhong, L., et al., Enhanced shot boundary detection using video text information. IEEE Transactions on Consumer Electronics, 2005. **51**(2): p. 580-588.*

*57. Wang, F., C.-W. Ngo, and T.-C. Pong, Structuring low-quality videotaped lectures for cross-reference browsing by video text analysis. Pattern Recognition, 2008. **41**(10): p. 3257-3269.*

58. Tuna, T., J. Subhlok, and S. Shah. Indexing and keyword search to ease navigation in lecture videos. in 2011 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). 2011.

59. Lu, T., et al., Video Text Detection. 2014: Springer Publishing Company, Incorporated. 264.

*60. Jung, K., K. In Kim, and A. K. Jain, Text information extraction in images and video: a survey. Pattern Recognition, 2004. **37**(5): p. 977-997.*

*61. Zhang, H., et al., Text extraction from natural scene image: A survey. Neurocomputing, 2013. **122**: p. 310-323.*

62. Chen, D. and J. Luettin, A Survey of Text Detection and Recognition in Images and Videos. 2000.

63. Lu, W.L., et al., Learning to Track and Identify Players from Broadcast Sports Videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013. ***35****(7): p. 1704-1716.*

64. Xian-Sheng, H., M. Tao, and H. Alan, eds. Online Multimedia Advertising: Techniques and Technologies. 2011, IGI Global: Hershey, PA, USA. 1-352.

65. Anagnostopoulos, C.N.E., et al., License Plate Recognition From Still Images and Video Sequences: A Survey. IEEE Transactions on Intelligent Transportation Systems, 2008. ***9****(3): p. 377-391.*

66. Reina, A.V., et al., Adaptive traffic road sign panels text extraction, in Proceedings of the 5th WSEAS International Conference on Signal Processing, Robotics and Automation. 2006, World Scientific and Engineering Academy and Society (WSEAS): Madrid, Spain. p. 295-300.

67. Liu, X. and J. Samarabandu, An edge-based text region extraction algorithm for indoor mobile robot navigation. IEEE International Conference Mechatronics and Automation, 2005, 2005. ***2****: p. 701-706 Vol. 2.*

68. Qixiang, Y. and D. David, Text Detection and Recognition in Imagery: A Survey. IEEE transaction on Pattern Analysis and Mahine Intelligence, 2015. ***37****(7): p. 1480 - 1500.*

69. Uchida, S., Text Localization and Recognition in Images and Video, in Handbook of Document Image Processing and Recognition, D. Doermann and K. Tombre, Editors. 2014, Springer London: London. p. 843-883.

70. Sonka, M., V. Hlavac, and R. Boyle, Image Processing, Analysis, and Machine Vision. 2007: Thomson-Engineering.

71. Jahne, B., Practical Handbook on Image Processing for Scientific and Technical Applications, Second Edition. 2004: CRC Press, Inc.

72. Nasrollahi, K. and T.B. Moeslund, Super-resolution: a comprehensive survey. Mach. Vision Appl., 2014. ***25****(6): p. 1423-1468.*

73. Rao, Y. and L. Chen, A survey of video enhancement techniques. Journal of Information Hiding and Multimedia Signal Processing, 2012. ***3****(1): p. 71-99.*

74. Poulose, M., Literature Survey on Image Deblurring Techniques. International Journal of Computer Applications Technology and Research, 2013. ***2****(3): p. 286-288.*

75. Guliashki, V. and D. Dimov. Image deblurring methods and image quality evaluation. in International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST'2014). 2014. Serbia.

76. Junker, M., R. Hoch, and A. Dengel. On the evaluation of document analysis components by recall, precision, and accuracy. in Document Analysis and Recognition, 1999. ICDAR '99. Proceedings of the Fifth International Conference on. 1999.

77. Yao, C., et al., Detecting texts of arbitrary orientations in natural images, in CVPR. 2012, IEEE Computer Society. p. 1083-1090.

78. Chen, H., et al., Robust Text Detection in Natural Images with Edge-enhanced Maximally Stable Extremal Regions, in 2011 IEEE International Conference on Image Processing. 2011: Brussels.

79. Il, K.H. and K.D. Hoon, Scene text detection via connected component clustering and nontext filtering. Image Processing, IEEE Transactions on, 2013. ***22****(6): p. 2296-2305.*

80. Pan, Y.-F., X. Hou, and C.-L. Liu, A Hybrid Approach to Detect and Localize Texts in Natural Scene Images. IEEE Transactions on Image Processing, 2011. ***20****(3): p. 800-813.*

81. Wang, H.-C., et al. Spatially Prioritized and Persistent Text Detection and Decoding. 2014. Cham: Springer International Publishing.

82. Rong, L., W. Suyu, and Z. Shi. A Two Level Algorithm for Text Detection in Natural Scene Images. in 2014 11th IAPR International Workshop on Document Analysis Systems. 2014.

83. Balagopalan, A., et al., Automatic keyphrase extraction and segmentation of video lectures. 2012 IEEE International Conference on Technology Enhanced Education (ICTEE), 2012: p. 1-10.

84. Wang, F., C.-W. Ngo, and T.-C. Pong, Structuring low-quality videotaped lectures for cross-reference browsing by video text analysis. Pattern Recogn., 2008. ***41****(10): p. 3257-3269.*

85. Dai, X., et al., Pseudo-Zernike moment invariants to blur degradation and similarity transformation. International Journal of Computer Mathematics, 2014. ***91****(11): p. 2403-2414.*

86. Tan, X., et al., Face recognition from a single image per person: A survey. Pattern Recogn., 2006. ***39****(9): p. 1725-1745.*

87. Gorji, H.T. and J. Haddadnia, A novel method for early diagnosis of Alzheimer's disease based on pseudo Zernike moment from structural MRI. Neuroscience, 2015. ***305****: p. 361-371.*

88. Jeong, H.J., et al., Automatic detection of slide transitions in lecture videos. Multimedia Tools and Applications, 2014. ***74****: p. 7537-7554.*

89. Powers, D.M., Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies, 2011. ***2****(1): p. 37-63.*

90. Lin, H.-Y., M.-Y. Hsieh, and K.-C. Li, Flexible group key management and secure data transmission in mobile device communications using elliptic curve Diffie-Hellman cryptographic system. Int. J. Comput. Sci. Eng., 2016. ***12****(1): p. 47-52.*

91. Merler, M. and J.R. Kender. Semantic keyword extraction via adaptive text binarization of unstructured unsourced video. in 2009 16th IEEE International Conference on Image Processing (ICIP). 2009.

92. Lu, H., et al., Image restoration using anisotropic multivariate shrinkage function in contourlet domain. International Journal of Computational Science and Engineering (IJCSE), 2016. ***12****(2/3): p. 95 - 103.*

93. Singh, C., et al., Analysis of algorithms for fast computation of pseudo Zernike moments and their numerical stability. Digital Signal Processing, 2012. ***22****(6): p. 1031-1043.*

94. Belkacem, S., L. Guezouli, and S. Zidat, Text Detection and Localization in Lecture Videos Using Moments, in Second International Conference on Internet of Things and Cloud Computing (ICC'2017). 2017: Algeria.