République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahlab Blida

Faculté des sciences

Département d'informatique

Mémoire Présenté par :
Mostefaoui Souad AICHOUCH Hadjer

En vue d'obtenir le diplôme de master

Domaine : Mathématique et informatique

Filière : Informatique
Spécialité : Informatique
Option : Génie des System Informatique

Sujet :

Génération de résumés vidéo

Soutenu le :

Mr Ouldkhaoua A	Président
Mr Chemchem A	Examinateur
Mr Kameche Abdelah Hicham	Promoteur
Mme Yahia Zoubir Bahiya	Encadreur

Promotion
2015/ 2016

Remerciement

Tout d'abords, nous remercions le dieu de nous avoir donné la santé et la volonté pour mener a bien ce modeste de travaille. Nous commençons d'abord par adresser notre sincères remerciements à madame Yahya Zobir Bahia et madame Ait Sadi Karima, pour nous avoir accordé l'honneur de travailler avec eux, pour nous' avoir permis d'élargir notre connaissance, et de mettre un pas dans le grand monde de la recherche et elles nous avons proposées des nouvelles orientations, en plus on les remercie également pour leurs sympathie, leurs modestie, leurs implication dans la réussite de ce travail. Nous exprimons notre profonde gratitude à Monsieur kamech Abdallah Hicham pour la qualité de ses conseils et pour les nombreuses discussions que nous avons vues. Nous tenons à remercier les membres de notre jury de nous avoir l'honneur d'être examinateurs tout en apportant leurs remarques et leurs contributions à l'enrichissement de ce mémoire. Nous remercions encore tous les enseignants du département .

Dédicaces

Je dédie ce modeste travail :

A la mémoire de mon défunt père

À la plus belle créature que dieu a créée sur terre....

À cette source de tendresse, de patience et de générosité A ma mère !

À mes chères sœurs Nadia, Hayet et Meriem, ainsi que leurs maries et leurs enfants

Djaber, Tarek et Anfel qui sont toujours de m'encourager.

À mon cher frère Toufik qu'il prenne le rôle et les responsabilités de père

À tout ma famille

À toute mes copines

À mes chère amies Asma et Hafida

A tout les étudiants GSI de promo 2015/2016.

À ma chère binôme Hadjer

M .Souad

Dédicaces

Je dédie ce modeste travail :

A la mémoire de mon frère

À la plus belle créature que dieu a créée sur terre....

MA très chère et douce mère, Mon très cher père à qui m'adresse au ciel les vœux les plus ardents pour la conservation de leur santé et de leur vie.

A celle qui ma appris le sens de la patience et celle qui nia jamais cessé de donner l'aide à chaque fois que j'en ai besoin, A ma chère mère !

À mes chères sœurs , ainsi que leurs enfants Loudjaine, Hiba,Sara et Aymen qui sont toujours de m'encourager.

À mon cher frère Hamza .

À mes chère amies Asma et Hafida.

À toute mes copines

A tout les étudiants GSI de promo 2015/2016.

À ma chère binôme Souad

A.Hadjer

Sommaire

Table des figures

Liste des tableaux

Résumé

Chaque jour des millions d'heures de séquences vidéo sont acquises et sauvegardées à travers le monde. Ces données vidéos sont tellement volumineuses qu'elles deviennent difficiles à traiter avec des techniques classiques de gestion de l'information. Dans ce mémoire, nous proposons une solution qui consiste à développer un algorithme de génération de résumés vidéos basé sur l'extraction des images clés qui va permettre de générer une version compacte de la vidéo initiale tout en préservant les informations d'intérêt. Pour le faire nous proposons deux méthodes : l'une basée sur les informations de bas niveaux et l'autre basée sur des informations de haut niveau afin d'extraire les trames représentatives d'une vidéo connues sous le nom de « Key-frames ».

Mots Clés : résumés vidéos, image clé, carte de saillance ,histogramme spatiale.

Abstract

Every day millions of hours of video are acquired and saved worldwide. These video data are so large that they become difficult to treat with conventional information management techniques. In this paper, as a solution to this problem, we propose video summaries generation algorithms based on the extraction of key frames that will allow generating a compact version of the original video while preserving the information of interest. To do so we propose two methods : one based on the low levels of information and the other based on the high level of information to extract the representative frames of a video known as the "key-frames" .

Keywords : video summaries, key-frames, saliency map ,spatiogram.

introduction générale

Chaque jour des millions d'heures de séquences vidéo sont acquises et sauvegardées à travers le monde. Ces données vidéo sont tellement volumineuses qu'elles deviennent difficiles à traiter avec des techniques classiques de gestion de base de données ou de gestion de l'information. Par conséquent, beaucoup des recherches ont été faites afin de réduire la quantité d'information qui doit être stockée ou traitée, une des solutions proposée est la génération de résumés du contenu vidéo. La génération de résumé vidéo est une technique alternative prometteuse utilisée dans l'indexation et la recherche vidéo. Dans ce mémoire nous proposons un nouvel algorithme de génération de résumés vidéo basé sur les cartes de saillances, cet algorithme va permettre de générer une version compacte de la vidéo initiale tout en préservant les informations d'intérêt, il va permettre aussi une manipulation aisée et une visualisation rapide du contenu des vidéos.

Afin d'aborder tout les aspects ayant trait à l'objectif de notre méthodologie, le mémoire est organisé comme suit :

- Chapitre 1 : une description de la structure, différents composants et caractéristiques d'une vidéo sont présentés suivi d'un état de l'art sur quelques méthodes de création de résumé de vidéo.
- Chapitre 2 : dans ce chapitre, nous nous intéresserons à la segmentation temporelle du contenu vidéo connue sous le nom de segmentation en plans
- Chapitre 3 : Dans ce chapitre nous nous intéressons aux différentes méthodes de modélisation de l'attention visuelle.
- Chapitre 4 : dans ce chapitre nous allons décrire la méthode que nous avons proposée pour l'extraction des images clés à partir d'une vidéo ainsi que la présentation de notre application et les résultats obtenus.

Chapitre 1 : état de l'art

Chapitre 1

état de l'art

1.1 Introduction

Avec le développement des nouvelles technologies, chacun d'entre nous possède au minimum un Smartphone doté de caméras numériques, avec lesquelles nous avons pris l'habitude de filmer les différents évènements de notre vie quotidienne dans le but de les partager via internet où de les immortaliser en les sauvegardant sur des supports de stockages tels que les disques durs, les cartes mémoires...etc. D'autres moyens de création, ou de sauvegarde de documents vidéo existent ou sont en plein essor, comme les documents vidéo créés à partir d'enregistrements satellites, les enregistrements des appareils médicaux ou des caméras de surveillance, ainsi que les flux multimédia diffusés par des milliers de chaînes télévisées, ou présents sur le Net. De plus, la technologie numérique permet la construction de vidéos en utilisant plusieurs images, ce qui induit l'augmentation du besoin de la sauvegarde de ces documents vidéo, mais il devient de plus en plus difficile de trouver le temps pour revoir ces documents et de la place pour les sauvegarder. Ces quantités énormes de données multimédia ont de loin dépassé la capacité que nous avons à toutes les traiter, et en tirer avantage.

Cette situation s'accentue au fil du temps, et le développement de nouveaux outils plus «intelligents» pour faire face à ce problème deviennent indispensables. La création automatique de résumés vidéos [116] est un outil performant qui permet de résumer le contenu général de la vidéo et de présenter les parties les plus pertinentes sous forme d'une séquence audiovisuelle ou d'un ensemble d'images représentatives. Les résumés vidéos permettent d'avoir rapidement une idée sur le contenu de très grandes bases de vidéos,

sans nécessiter la visualisation et l'interprétation de l'ensemble des vidéos. Cela permet aussi de juger et d'évaluer la pertinence d'un document multimédia par rapport aux autres.

Dans ce chapitre, une description de la structure et des différents composants et caractéristiques d'une vidéo sont présentés suivi d'un état de l'art sur quelques méthodes de création de résumé de vidéo.

1.2 Notions générales sur la vidéo numérique

Une vidéo se compose d'images affichées à une fréquence de 25 images (ou 30 images) par seconde, accompagnées d'une bande son. Contient différentes élément et caractéristique.

1.2.1 Définition d'une vidéo numérique

Un document vidéo peut être vu comme étant la combinaison de deux modes : la vidéo et l'audio, représentés dans un espace de temps discret associé, en général, à une fréquence d'échantillonnage plus élevée que les changements d'états qu'il reflète, ce qui permet au spectateur de le percevoir comme étant continu.

1.2.2 Structure d'une vidéo

Une navigation aisée à l'intérieur d'un document vidéo nécessite que son contenu soit étayé par une structure non-linéairement dépendante du temps [71]. Ils s'agit notamment de permettre à un utilisateur de ne regarder qu'un thème sélectionné redondant dans la vidéo (il doit par exemple pouvoir choisir de ne visionner que des scènes d'extérieur dans un documentaire),la Figure 2.2 illustre les différents éléments d'une vidéo.

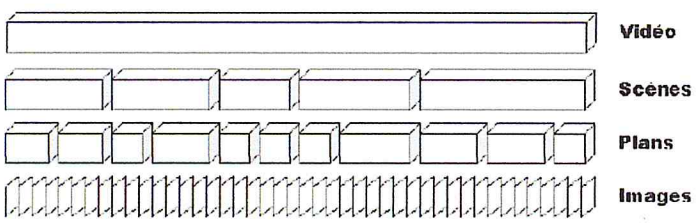


FIGURE 1.1 – Structure d'une vidéo

1.2.3 Les éléments d'une vidéo

Une vidéo peut être décomposée en plusieurs éléments :

Coupe (shot)

elle est définie comme étant une transition immédiate d'une scène à l'autre qui se produit entre deux plans [89]. Par définition, une transition correspond au point de jonction entre deux plans. Il existe plusieurs types de transitions dans les vidéos. Celles-ci ont été regroupées suivant deux grandes familles de transitions :

- *Les changements progressifs* : qui consistent en l'obtention d'une continuité visuelle lors du passage d'un plan à l'autre. Cette transition est réalisée soit par fondu enchaîné(figure 2.3) [32], soit par changement progressif (figure 1.3)[89] de la couleur de la séquence jusqu'à atteindre une teinte uniforme (fade in/fade out) [74](figure 1.4).
- *Les changements de plans brusques (ou instantanée)* : qui consistent à juxtaposer la fin d'un plan avec le début du plan suivant sans transition [89]. Ces changements sont les plus utilisés dans les vidéos(figure 1.5).

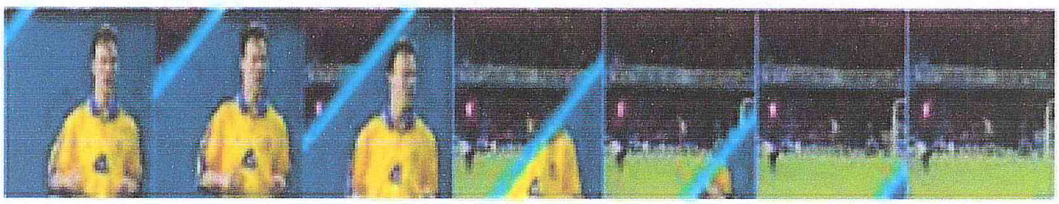


FIGURE 1.2 – Changements progressifs

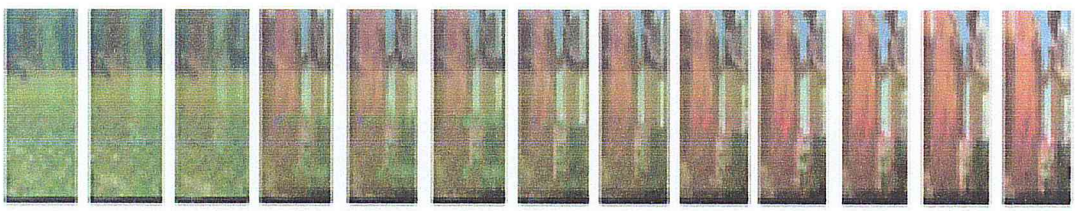


FIGURE 1.3 – Changements progressifs (Changements progressifs (fondu enchaîné))

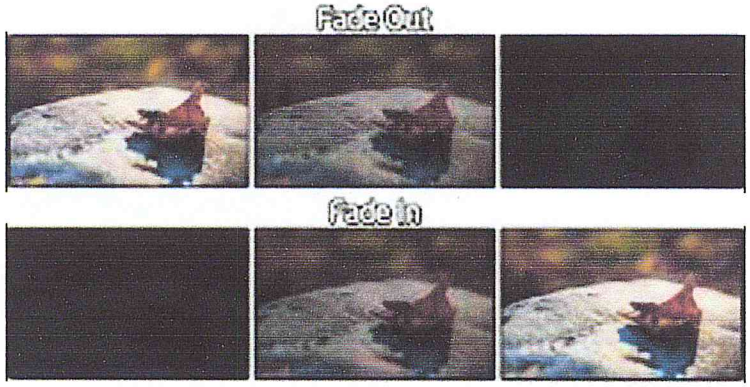


FIGURE 1.4 – Changements progressifs (fade in/fade out)

FIGURE 1.5 – Changements de plans brusques

Beaucoup de techniques ont été développées pour détecter les différents types de transitions. La plupart d'entre elles utilisent des différences entre images contenues dans une vidéo basées sur des comparaisons de luminances ou d'histogrammes [62]. A partir de caractéristiques visuelles ou temporelle (couleur, contour ou mouvement) extraites sur chaque image, une distance est définie soit entre deux images successives soit entre images contenues dans une fenêtre temporelle pour déterminer des différences entre elles [100]. De fortes variations sur les différences renseignent sur la présence d'une discontinuité temporelle et donc d'une possible transition [89]. Suivant la nature de la transition, instantanée ou progressive, différentes approches ont été proposées. Pour les transitions instantanées, la méthode la plus simple pour les détecter consiste à comparer les pixels entre images successives [117]. Cependant cette comparaison n'est pas insensible aux mouvements des objets et de la caméra, ce qui peut aboutir à de fausses détections. De ce fait, certaines approches développent des descripteurs moins sensibles aux mouvements comme les histogrammes [124]. D'autres utilisent la compensation de mouvement ou le suivi de caractéristiques pour créer une métrique entre les images [114]. Enfin, des comparaisons entre histogrammes couleur locaux (par division de l'image en blocs) peuvent aussi être effectuées pour être moins sensibles aux mouvements des objets [100]. En ce qui concerne les transitions progressives, leur détection est plus difficile en raison de la faible différence

entre deux images successives et le nombre inconnu d'images dans la transition. Pour être indépendantes de la longueur des transitions, les distances sont le plus souvent calculées à l'intérieur d'une fenêtre temporelle [97]. De plus, comme le mouvement de caméra entraîne souvent de fausses détections [96], certaines approches l'estiment pour distinguer les changements progressifs (liés au mouvement de la caméra) des transitions progressives [100]. D'autres caractérisent les transitions à partir de la prédiction de blocs (utilisation du bloc matching) [12]. Par ailleurs, une étape importante dans la détection est la détermination des seuils. La plupart des méthodes utilisent des seuils préfixés [89]. D'autres approches définissent des seuils dynamiques dépendants du contenu des images [114].

Une scène

Une scène est composée d'une suite de plans et possède les caractéristiques suivantes : unité de temps et unité d'action [89]. Une scène décrit un événement de manière continue [58]. Ainsi, un flash-back est considéré comme une nouvelle scène. Il est de même lorsqu'un changement de décor intervient entre deux plans, indiquant qu'un laps de temps significatif a passé. Une scène décrit un seul événement. Même en cas de continuité temporelle, il peut arriver qu'un nouveau personnage ou un événement extérieur vienne perturber le déroulement de l'histoire : dans ce cas, nous considérons qu'une nouvelle scène commence [60].

Un plan

le plan est une unité de montage [33]. C'est une prise de vue de la caméra, un morceau d'enregistrement effectué sans interruption ayant un début et une fin, il se compose d'une limite, le cadre du plan est la surface visible de l'image [5].

- *La construction du plan* : le cadre met en valeur certains éléments, selon la composition de l'image (amorce de plan, premier plan, second plan, arrière-plan) [19]. La profondeur de champ dans la composition du plan est utilisée pour mettre en valeur les éléments que l'on souhaite (figure 1.7)

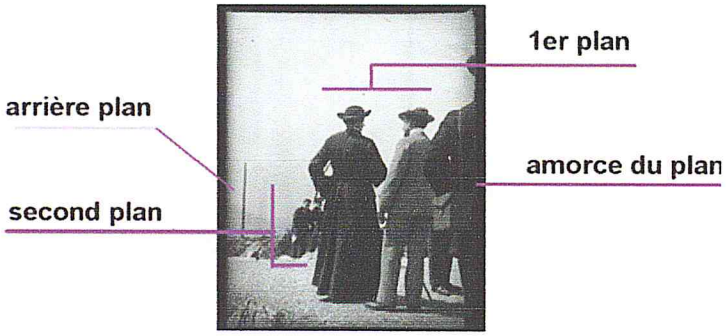


FIGURE 1.6 – représentation d'un plan

Il existe différents types de plans représentés dans la figure suivante :

FIGURE 1.7 – présente les types de plan

L'image

une image numérique (2D ou 3D) est l'unité de base d'une vidéo, est une matrice $n \times m$ (lignes et m colonnes) définie par sa résolution (taille du pixel "Picture élément"), sa profondeur (8 bits, 16 bits...etc), et sa table de couleur (color look up table (CLUT)) [26].

Une image numérique est composée d'un ensemble fini d'éléments, appelés Picture élément, ou pixels. Le pixel est un élément fini, uniformément coloré, et est le plus petit d'une image [1](figure 1.8).

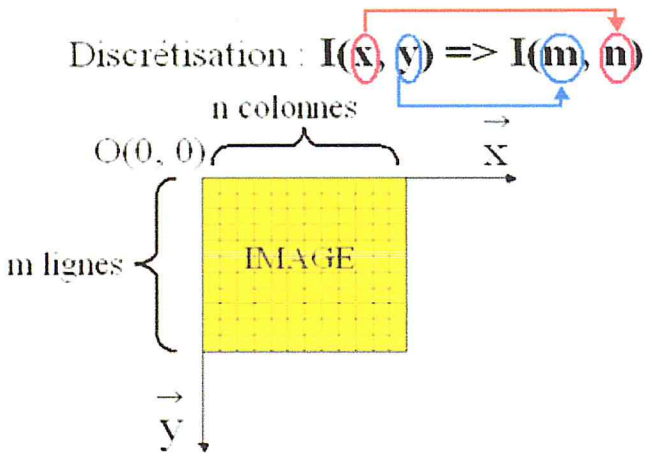


FIGURE 1.8 – représentation d'une image numérique 2D

- ***Structure d'une image*** : Une image est la représentation d'un être ou d'un objet obtenue par la photographie, la vidéo ou l'utilisation d'un logiciel spécialisé. Elle est dite numérique lorsque sa sauvegarde est obtenue sous forme binaire. Donc image numérique fait appel à l'informatique [17].
  Chaque image numérique est constituée d'un nombre donné de lignes. Chaque ligne comporte un nombre de point donnés. L'ensemble constitue une matrice. Ces points sont dénommés pixel (de l'anglais Picture élément). Chaque « case » de cette matrice contient des nombres caractéristiques à la couleur attribuée au pixel [21].

- ***Résolution :*** La résolution d'une image est définie par le nombre de pixels par unité de longueur [34]. Usuellement, le nombre de pixels est compté par pouce (1 pouce = 2,54 cm, noté ppp ou dpi) ou par centimètre [39]. Plus le nombre de pixels par unité de longueur est élevé, plus la quantité d'information décrivant l'objet est importante donc la résolution est grande. Son corolaire est une taille de fichier de plus en plus importante [17]. Ce paramètre est défini souvent lors de l'acquisition de l'image (réglage de l'appareil photo, résolution du logiciel du scanner... etc.) ou ultérieurement dans les logiciels de traitement d'images. La publication d'image sur Internet correspond souvent à une résolution de 90 *ppp* (points par pouce : un pouce = 25,4 mm) et dans la presse écrite de 150 *ppp* [51].

- ***Taille de l'image :*** La résolution de l'image influe directement sur la taille du fichier de sauvegarde de celle-ci. Par exemple, dans le cadre de la télévision, nous

avons :

la télévision à définition standard SD 576 a 768 pixels par ligne et 576 lignes par image soit un total de 442368 pixels par image [55].

La télévision haute définition HD 1080 a 1920 pixels par ligne et 1080 lignes par image soit un total de 2073600 pixels par image [52]. Ainsi si la résolution est environ multipliée par 2, les nombre de pixels l'est par 4 (structure bidimensionnelle de l'image) ce qui peut engendrer un temps de traitement plus long. Il faut donc trouver un compromis entre la qualité attendue de l'image en termes de résolution et la taille de son fichier de sauvegarde.

- ***Type d'image*** :Il existe deux types d'image, l'image matricielles et l'image vectorielles.

  Images matricielles : Dans la description des images que nous avons faite jusqu'à présent nous avons utilisé une matrice. Nous avons dit alors que l'image est matricielle ou en anglais bitmap [59]. Ce type d'image est adapté à l'affichage sur écran mais peu adapté pour l'impression car bien souvent la résolution est faible (couramment de 72 à 150 ppp pour les images sur Internet).

  Images vectorielles : Le principe des images vectorielles [64] est de représenter les données de l'image à l'aide de formules mathématiques. Cela permet alors d'agrandir l'image indéfiniment sans perte de qualité et d'obtenir un faible encombrement.

1.2.4 Caractéristiques d'une vidéo

La vidéo est caractérisée par la couleur, le mouvement et le texture :

Représentation de la couleur

La couleur est une notion complexe et sa perception par l'homme a fait l'objet de nombreux travaux de recherche depuis ceux de Newton il y a plus de trois siècles [5]. Le codage de la couleur dans des images numériques peut être effectué en utilisant différents espaces de représentation, appelés traditionnellement espaces couleur. Nous présentons ici les principaux espaces définis dans la littérature.

- ***Espaces proposés par la CIE*** :

L'espace RGB proposé par la CIE (Commission Internationale de l'Éclairage) est très certainement le système de représentation couleur le plus fréquemment utilisé [30]. Chaque couleur est définie par une combinaison de trois couleurs primaires : Rouge (R), Vert (G), et Bleu (B). Dans cet espace, le noir et le blanc sont représentés respectivement par les triplets ($R = 0, G = 0, B = 0$) et ($R = 255, G = 255, B = 255$) si l'on considère le cas d'images numériques couleur codées sur 24 bits. Il est important de noter que d'autres espaces de représentation RGB ont été définis, notamment par la Fédéral Communication Commission (FCC) pour le standard de télévision $NTSC$ (National Télévision Standards Committee) ou par l'European Broadcasting Union (EBU) pour le standard PAL (Phase Alternation by Line) [62]. Un processus de normalisation des composantes du vecteur couleur permet de réduire l'influence de la luminance [98]. L'espace couleur RGB normalisé est noté rgb. Dans cet espace, les différentes composantes ne représentent plus qu'une information de chrominance car les composantes sont liées par la contrainte $r + g + b = 1$. La CIE a aussi défini un autre espace de représentation basé sur des primaires. Il s'agit de l'espace XYZ dans lequel la composante Y représente la luminance et les deux autres composantes des informations additionnelles relatives à la chrominance. De même qu'avec l'espace RGB, il est possible de définir un espace XYZ normalisé. Cet espace est noté xyz avec $x + y + z = 1$.

- ***Espaces luminance-chrominance*** **:**

En se basant sur les espaces XYZ nous pouvons obtenir l'espace YUV généralement utilisé dans les codages des séquences au format $MPEG$ [10]. La composante Y est commune à l'espace XYZ et représente donc toujours la luminance, tandis que les composantes U et V contiennent les informations de chrominance. La CIE a défini également des espaces appelés $L * u * v$ et $L * a * b$ fréquemment utilisés dans la littérature [110]. La composante L représente ici la luminance tandis que les couples (u, v) et (a, b) permettent de modéliser la chrominance.

- ***Espaces dans le domaine de l'imprimerie*** **:**

Dans le domaine de l'imprimerie, l'espace CMY est souvent utilisé [49]. Cet espace, "complémentaire" de RGB, considère les trois couleurs Cyan (C), Magenta (M),

et Jaune (Y), calculées comme les complémentaires de Rouge, Vert, et Bleu. Il est aussi possible d'utiliser une représentation quadri chromatique $CMYK$, où la composante additionnelle K représente le noir [110].

- ***Espaces dans le domaine télévisuel*** :
  Pour les usages du monde télévisuel, plusieurs systèmes ont été définis. Le standard $NTSC$ est basé sur le système $Y_0I_0Q_0$de la FCC [63]. Le standard PAL utilise quant à lui le système $Y_0U_0V_0$ de l'EBU, relativement similaire au précédent [85] . Kodak a aussi proposé son propre système, appelé YCC ou YC_1C_2.

Description du mouvement

Le mouvement est une information riche qui renseigne sur l'activité d'un plan et celle de ses objets. A partir de la séquence des images formant un plan, le mouvement de la caméra peut être estimé ; ensuite le mouvement des objets peut être déterminé [73]. L'ensemble des vecteurs de mouvement des points ou des régions est communément appelé flux optique [20]. Il est à la base de tous les systèmes d'analyse du mouvement des scènes. De nombreuses techniques de calcul de flux optique ont été développées à partir des années 80 et de nouvelles techniques sont encore proposées de nos jours [87]. Le flux optique est employé dans des domaines très variés comme l'imagerie médicale, la robotique, la télésurveillance, la compression et l'indexation, qui ont leurs propres contraintes [102]. Sans être exhaustif, il permet d'effectuer la détection et le suivi d'objet, de modéliser l'environnement par reconstruction 3D, d'estimer le mouvement de la caméra, d'effectuer une segmentation spatio-temporel en région.

La texture

La texture est une information de plus en plus utilisée en indexation d'images et de la vidéo. Elle permet de combler un vide que la couleur est incapable de faire, notamment lorsque les distributions de couleur sont très proches. La texture est une région de l'image qui a des caractéristiques cohérentes et homogènes, formant un tout pour un observateur. Elle est composée de petits éléments répétitifs. Généralement, la répétition peut impliquer des variations locales d'échelle, d'orientation, ou d'autres caractéristiques géométriques et optiques des éléments.

1.3 Notions générales sur la génération de résumés vidéos

Devant le volume grandissant des données audiovisuelles, la construction automatique de résumé vidéo [32]est devenue un domaine de recherche en pleine expansion. Le résumé vidéo est défini comme étant l'ensemble des images clés (key frame) [68] de cette vidéo. Ces images clés sont définies comme les images les plus informatives qui capturent les principaux éléments d'une vidéo en termes de contenu, ce qui permet de localiser rapidement les informations les plus pertinentes présentent dans la vidéo. Le résumé de vidéo a pour objectif de fournir des informations pertinentes et concises afin d'aider l'utilisateur à naviguer ou à organiser ses fichiers vidéos plus efficacement et en utilisant un espace de stockage réduit. D'autre part la construction de résumés vidéos fait gagner aux utilisateurs un temps considérable et contribue au développement de techniques qui seront utilisées pour des applications nouvelles comme par exemple la télévision interactive [67].

Différents travaux de recherche ont traité le problème de construction des résumés vidéos en apportant diverses solutions et propositions, malgré que ce soit un domaine de recherche assez récent mais en plein essor, ces travaux peuvent être classé selon la méthode avec laquelle le résumé vidéo a été extrait, nous pouvons citer entre autres :

- Méthodes basées sur l'histogramme de couleurs.
- Méthodes basées sur les régions clés.
- Méthodes basées sur le mouvement.
- Méthodes basées sur l'entropie.
- Méthode basées sur le découpage en segments.
- Autres méthodes qui sont obtenues en combinant les méthodes précédemment citées, ou qui utilisent de nouvelles approches.

1.3.1 Algorithmes basés sur l'histogramme de couleurs

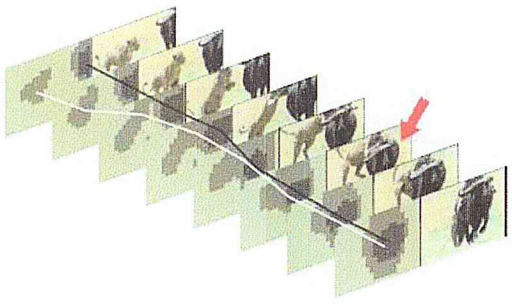
L'histogramme de couleurs est invariant aux orientations de l'image et robuste au bruit. C'est pourquoi, les algorithmes d'extraction d'image clés fondés sur la couleur ont été fortement exploités.

Les travaux de Y. Zhang [121] proposent d'utiliser la couleur pour détecter des images clés. Une fois que le découpage en plans a été réalisé, la première image est systémati-

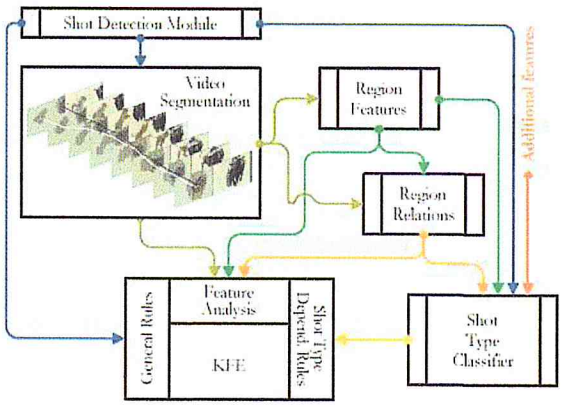
quement déclarée comme une image-clé et devient l'image de référence. L'histogramme de cette image de référence est ensuite comparé aux histogrammes des images suivantes. Le parcours du plan est effectué chronologiquement. Lorsque la distance entre l'histogramme de la référence et celui d'une image donnée excède un certain seuil, l'image courante est déclarée comme la nouvelle référence et devient par la même occasion une image-clé. Dans [125], Huang et al. Proposent une méthode de classification (ou clustering) fondée sur la similarité des histogrammes couleur des images appartenant à un même plan. Les images-clés correspondent aux images les plus proches des centroïdes des différentes classes (ou clusters). L'inconvénient de la plupart de ces travaux réside dans le fait qu'ils sont extrêmement dépendants d'un seuil puisqu'ils découlent de la comparaison d'histogrammes.

1.3.2 Algorithmes basés sur les régions clés

Dans [13], J. Calic et al. proposent une étude temporelle du comportement de régions clés obtenues par une segmentation spatiale à basse résolution. Cette segmentation est réalisée par classification des coefficients de la DCT (Discrete Cosine Transform) directement issus d'un signal vidéo compressé. La figure 1.9.a présente une segmentation en régions d'une suite d'images où il est possible de visualiser les trajectoires des deux régions d'intérêt. Les images-clés sont sélectionnées suivant des règles fondées sur les disparitions, les apparitions et les interactions entre les régions. Certaines règles sont générales, tandis que d'autres sont inhérentes au type de plan détecté. La figure 1.9.b présente les différentes étapes de la méthode.



(a) Images originales et leur segmentation avec visualisation des trajectoires des régions



(b) Synopsis de de la méthode

FIGURE 1.9 – Extraction d'images-clés par suivi de régions-clés (d'après [13])

1.3.3 Algorithmes basés sur le mouvement

Les approches orientées mouvement ont l'avantage d'adapter le nombre d'images-clés à la dynamique temporelle de la scène. Les méthodes les plus utilisées sont les méthodes de différence d'images [57] ou de flux optique [115].
Dans [115], W. Wolf propose de calculer pour chaque image, une mesure simple du mouvement à partir du flux optique. Cette mesure M correspond à la somme des amplitudes du flux optique des N pixels p_i de l'image traitée ($\forall n \in J1, NK$). M est donné par l'équation suivante :

$$M = \sum_{i=1}^{N} \mid o_x(p-i) \mid + \mid o_y(p_i) \mid \tag{1.1}$$

Où o_x (p_i) et o_y (p_i) sont respectivement l'abscisse et l'ordonnée du flux optique calculés en le p_i ème pixel de l'image courante. Par analyse des variations temporelles de cette mesure au cours du plan, les images correspondant aux minima locaux de mouvement sont sélectionnées pour devenir les images-clés. Selon l'auteur, l'analyse en mouvement permet de découper un plan en plusieurs évènements significatifs. Le mouvement est considéré comme un outil que le réalisateur peut utiliser pour mettre en valeur certaines parties du plan. En l'occurrence des mouvements lents sont considérés ici, comme synonymes d'un événement important. Dans le cas d'applications plus spécialisées, Ju et al. Réalisent la création d'un résumé dans le contexte d'une présentation d'un exposé [48]. L'extraction des images-clés est fondée sur une reconnaissance du mouvement de la caméra et des gestes des personnes.

1.3.4 Algorithmes basés sur l'entropie

Dans [72], Markos Mentzelopoulos et al. Proposent une méthode d'extraction d'images-clés fondée sur la différence d'entropie entre les images. Le principe de l'algorithme est d'isoler dans chaque image les zones saillantes qui présentent une entropie supérieure à 70 % de l'entropie totale de l'image. Théoriquement, ces zones sont supposées être de forte représentativité puisqu'elles contiennent 70 % de l'information disponible dans l'image. L'extraction des images-clés est fondée sur l'étude de la différence entre deux images successives de l'entropie contenues dans les zones de forte représentativité. Lorsque la différence excède un certains seuil, une nouvelle image-clé est déclarée. Selon les auteurs,

l'algorithme gère bien les images où les objets et les personnages se détachent correctement du fond. Cependant, les performances sont fortement dégradées lorsque la séquence contient des flashes, comme des explosions par exemple.

1.3.5 Algorithmes basés sur le découpage en segments

Un des inconvénients majeurs dans l'utilisation d'une ou plusieurs images-clés pour chaque plan est l'incompatibilité avec les longues vidéos comportant un nombre important de plans. C'est pourquoi de récentes recherches se focalisent sur l'aspect haut-niveau des vidéos et tentent d'extraire cette fois des segments de la vidéo qui peuvent être soit une scène soit un événement particulier voire la séquence entière. L'ensemble des images-clés sélectionné par ce type de méthodes est plus concis que celui issu de plans. Le système vidéo manga [107] proposé par Uchihashi et al., crée un résumé organisé comme une bande dessinée (figure 1.10) . Leur approche consiste à effectuer une classification de l'ensemble des images de la vidéo en utilisant la similarité de leurs histogrammes couleur dans l'espace YUV (est le standard Européen PAL [1], et $SECAM$ [2] de transmission des images couleur pour la télévision). De cette classification résulte une segmentation de la vidéo indépendante d'un découpage en plans classique. En effet, chaque segment est déterminé par les images contiguës appartenant à la même classe. Ensuite une mesure de l'importance de chaque segment est évaluée par rapport à sa taille et sa rareté. Un segment est d'autant moins important qu'il est court et similaire aux autres segments. Cette mesure permet de supprimer des segments non significatifs et de hiérarchiser ceux restants. Les images-clés présentes dans le résumé sont les images situées au centre de chaque segment retenu. Leur taille est proportionnelle à l'importance des segments dont elles sont extraites.

1.3.6 Autres algorithmes

Les paragraphes précédents ne représentent qu'un aperçu des méthodes principales présentes dans la littérature ; ce domaine de recherche étant tellement actif. Ainsi, il existe de nombreuses méthodes d'extraction d'images-clés utilisant des outils différents et plus spécialisés telles que la détection de visage ou la transformée en ondelettes. Cependant,

1. Phase Alternative Line
2. Séquentiel Couleur Avec Mémoire

FIGURE 1.10 – Exemple d'un résumé de style Bande Dessinée [107]

il existe une certaine catégorie d'algorithmes qui concerne davantage notre travail de fin d'étude, c'est la construction de résumés fondés sur les objets.

1.4 Conclusion

Dans ce chapitre, nous avons présenté quelques notions de base sur la vidéo, ainsi qu'une revue générale sur quelques approches d'extraction de résumés vidéos qui reposent sur des caractéristiques différentes. Le but de ce chapitre est de donner une vue d'ensemble de ce qui se fait dans le monde scientifique qui s'intéresse à la génération de résumés vidéos afin de situer notre travail par rapport à eux dans le chapitre suivant. Nous avons observé que l'extraction des résumés vidéo est un thème de recherche très vaste qui nécessite la mise en œuvre de nouveaux outils plus performants qui répondent au mieux aux exigences des sciences et technologies nouvelles.

Chapitre 2 :Notions de base sur la segmentation de la vidéo en plans (Shot boundary detection)

Chapitre 2

Notions de base sur la segmentation de la vidéo en plans (Shot boundary detection)

2.1 Introduction

Dans le chapitre précédant, nous avons donné quelques notions de base sur la vidéo et l'image, dans ce chapitre nous nous intéressant à la segmentation temporelle du contenu vidéo connu sous le nom de segmentation en plans. Elle est usuellement obtenue en détectant les changements de plans dans le flux vidéo. L'extraction de plans à partir d'une vidéo est un problème qui a été abondamment traité dans la littérature. Nous allons présenter dans ce chapitre un état de l'art des différentes méthodes proposées. La présentation n'a pas pour but d'être exhaustive mais plutôt d'être la plus variée possible. Détecter les plans à partir d'une vidéo permet d'obtenir des informations importantes sur sa structure. Ces informations seront utilisées dans les prochains chapitres, pour l'extraction des images clés.

2.2 Segmentation de la vidéo en plan

Pour rendre l'analyse et l'indexation de la vidéo plus facile, cette dernière doit être découpée en unités logiques telles que des plans et des scènes. En général, chaque plan est l'effet du mouvement d'une seule caméra [19]. Comme les images dans un plan sont

fortement corrélées le long de la dimension temporelle, très peu d'images représentatives ou images clés sont choisies pour récapituler chaque plan. Cependant, avec cette approche, une vidéo de longueur typique sera représentée par des milliers d'images clés. Par conséquent, les chercheurs ont proposé des méthodes pour identifier les images sémantiquement importantes et modéliser les rapports inter plans. Comme illustré dans la figure 2.1, le partitionnement d'une vidéo peut se faire à 4 niveaux différents de granularité [91] :

- Niveau-Image : chaque image est traitée séparément. Aucune analyse temporelle à ce niveau.
- Niveau-Plan : un plan est un ensemble d'images contiguës, toutes acquises par un enregistrement continu de la caméra.
- Niveau-Scène : une scène est un ensemble de plans contigus ayant une signification sémantique commune.
- Niveau-Vidéo : la vidéo est traitée comme un seul ensemble.

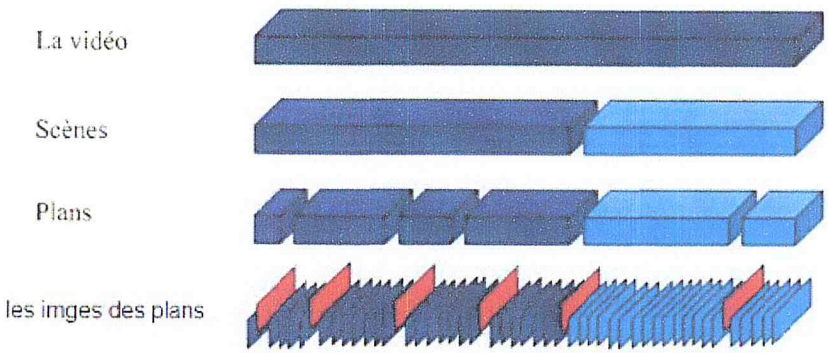


FIGURE 2.1 – Structure d'une vidéo.

Pendant le montage d'un document vidéo, le réalisateur inclut des transitions entre les différents plans et scènes. Dans ce qui suit, nous allons éclaircir le concept de transition et étudier les différentes techniques pour les détecter afin de découper les vidéos en plans.

2.2.1 Les types de Transition entre les plans

Les transitions entre les plans sont les moments de passage entre les plans de la caméra et qui mènent le lecteur de la vidéo à partir d'un plan à l'autre. Ils sont ajoutés pendant la postproduction. Il y a deux types de transitions que nous pouvons avoir entre les plans :

- les transitions de plan brusque (discontinu), également désignées sous le nom de coupure ou transition brusque [53].

- les transitions de plan progressif (continu), qui peuvent être du type estompé (fades), de dissolution (dissolve), de balayage (wipe) [119] .Ces transitions sont définies comme suit :

  1. Coupure (shot cut) : un changement brusque d'un plan à un autre (Figure 2.2(a)).
  2. Apparaître en fondu (fade-in) : le plan apparaît graduellement à partir de sa première image (Figure 2.2(b)).
  3. Disparaître en fondu (fade-out) : le plan disparaît graduellement (Figure 2.2(c)).
  4. Dissolution (dissolve) : le plan courant disparaît graduellement tandis que le prochain plan apparaît graduellement (Figure 2.2(d)).
  5. Balayage (wipe) : le prochain plan est indiqué par une frontière mobile sous forme d'une ligne ou d'un motif (Figure 2.2(e)).

(a) L'effet coupure (*shot cut*)

(b) L'effet apparaître en fondu (*fade-in*)

(c) L'effet disparaître en fondu (*fade-out*)

(d) L'effet dissolution (*dissolve*)

(e) L'effet essuyer (*wipe*)

FIGURE 2.2 – Différents types de transitions de plan.

La détection de l'une de ces transitions pendant l'analyse de la vidéo permet son découpage en plans. Dans la Figure 2.3, une brève description des différentes approches

existantes pour la détection des transitions brusques et graduelles entre les plans est donnée.

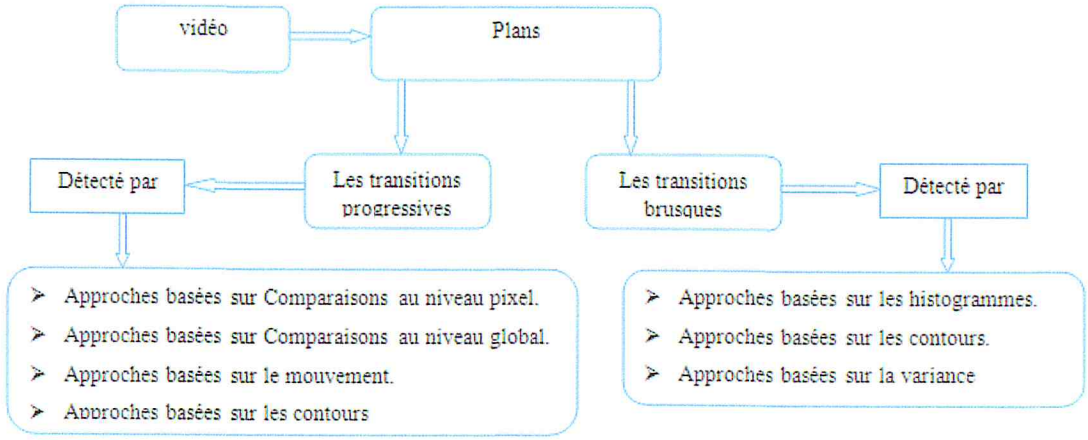


FIGURE 2.3 – différentes techniques de détection des plans.

2.3 La détection de transition brusque entre les plans

Cette section résume les approches existantes pour la détection des transitions brusques entre les plans. Au moment de la transition, une image est remplacée par une autre. Par conséquent, la détection doit produire un certain nombre de plans tels que :

- toutes les images dans le même plan exposent des caractéristiques semblables.
- les images appartenant à différents plans ont des caractéristiques différentes.

Différentes caractéristiques et métriques ont été proposées pour la détection de coupure entre les plans. Elles ont été analysées dans plusieurs études comparatives [11][70][65][118]. Dans ce qui suit, nous allons présenter les différentes approches utilisées pour la détection de coupure entre les plans :

1. Approches basées sur les comparaisons au niveau pixel.
2. Approches basées sur les comparaisons au niveau global.
3. Approches basées sur le mouvement.
4. Approches basées sur les contours.

2.3.1 Comparaisons au niveau pixel

La manière la plus simple de mesurer la différence entre deux images est de comparer les intensités des pixels entre les deux images. Selon la méthode proposée par [54], il y a une coupure entre deux images qui se succèdent dans une vidéo, si le changement de la moyenne des intensités des pixels est supérieur à un seuil donné. L'inconvénient de cette méthode est sa sensibilité aux mouvements des objets et de la caméra, car en utilisant le changement de la moyenne, il est impossible de faire la différence entre un grand changement dans une petite région de l'image et un petit changement dans une grande région. Zhang et al [122] ont proposé une amélioration qui consiste à déterminer le pourcentage des pixels qui ont changé considérablement entre deux images. Leur méthode utilise un filtre moyen 3x3 pour réduire le bruit et l'effet du mouvement de la caméra. Bien qu'elle apporte une amélioration, cette méthode est toujours sensible au mouvement des objets et de la caméra.

2.3.2 Comparaisons au niveau global

Quelques méthodes ont été proposées pour pallier au problème du mouvement de la caméra et des objets. Ces méthodes comparent les caractéristiques globales de chaque image au lieu de comparer chaque pixel individuellement. Nagasaka et Tanaka [75] ont proposé l'utilisation de l'histogramme à niveau de gris pour comparer deux images. Toutefois, la méthode n'était pas robuste en présence de bruit momentané, comme le flashe d'un appareil photo ou le mouvement d'un grand objet. Nagasaka et Tanaka [90] ont également proposé une autre méthode basée sur la comparaison de l'histogramme de la couleur. Ils ont proposé d'utiliser un code couleurs de 6 bits obtenus en prenant les deux bits les plus significatifs de chaque composante RGB ce qui donne un code à 64 couleurs. Ils utilisent la loi de Chi deux X2 [95] pour mesurer la différence entre deux distributions liées. Selon Gargi et al. [29], Nagasaka et Tanaka [75] et Lienhart [66], une simple comparaison entre les histogrammes de la couleur (RGB ou YUV), avec chaque bande quantifiée à 2^b (b nombre de bande) valeurs différentes, est une méthode efficace pour détecter les frontières des plans.

2.3.3 Comparaisons basées sur les blocs

Une faiblesse des comparaisons de niveau global est qu'elles peuvent manquer la coupure entre deux plans où seule la distribution spatiale du contenu change. Zhang et al. [75] proposent de partitionner l'image en régions (blocs) puis faire la comparaison des blocs correspondants dans deux images successives. Les blocs sont comparés sur la base des caractéristiques statistiques du deuxième ordre de leurs valeurs d'intensité. Nagasaka et Tanaka [75] proposent également de diviser chaque image en quatre régions et de comparer les histogrammes de couleur des régions correspondantes. Ueda et al. [108] proposent une autre approche en augmentant le nombre de blocs à 48 et en déterminant la mesure de différence entre deux images comme le nombre total de blocs avec une différence d'histogramme supérieur à un seuil donné. Selon Otsuj et al. [40], la méthode d'Ueda et al. [108] est plus efficace que celle de Nagasaka et Tanaka [75]. Cependant, le fait que les blocs soient plus petits dans l'approche d'Ueda et al. [108] signifie aussi que cette méthode est plus sensible au mouvement des objets et de la caméra [35]. Cela met en évidence le problème de choisir une échelle appropriée pour la comparaison entre les contenus de deux images.

2.3.4 Approches basées sur le mouvement

Plusieurs méthodes ont tenté d'éliminer la différence entre deux images causées par le mouvement des objets et de la caméra avant de faire la comparaison. Plusieurs auteurs [70][2][99] ont proposé des méthodes qui font la comparaison entre les images partitionnées en blocs afin d'obtenir des mesures de similarités basées sur le mouvement. La différence principale entre ces approches est la méthode utilisée pour combiner les mesures des blocs afin d'obtenir une caractéristique globale de l'image. Vlachos [113] a utilisé une méthode qui emploie la corrélation de phase afin d'obtenir une mesure de similarité entre deux images. Cette méthode se caractérise par son invariance aux changements dans l'illumination globale du contenu de l'image. Fernando et al. [24] ont exploité le fait que les vecteurs de mouvement sont de nature aléatoire pendant une coupure de plan. La méthode calcule le vecteur de mouvement moyen entre deux images et la distance Euclidienne par rapport au vecteur moyen pour tous les vecteurs de mouvement. Ils déduisent qu'il y a une coupure s'il y a une grande augmentation dans la distance Euclidienne.

2.3.5 Approches basées sur les contours

Zabih et al. [120] ont proposé une méthode pour détecter les transitions brusques de plan en vérifiant la distribution spatiale des contours sortants et entrants. Cette méthode a exploité le fait que les contours d'objets dans l'image avant une coupure de plan ne peuvent être trouvés dans le même emplacement après la coupure. Bien que cette méthode ait illustré la viabilité de la caractéristique de contour pour détecter un changement de la décomposition spatiale entre deux images, sa performance était décevante comparée avec d'autres métriques plus simples qui sont moins gourmandes en ressource machine [41][28][69].

2.4 La détection des transitions progressives

Contrairement aux coupures franches, la différence entre les images pendant une transition progressive est petite. Pour cette raison, il peut être difficile de distinguer entre les changements provoqués par le mouvement de la caméra et des objets et ceux provoqués par une transition progressive. Par conséquent, si on essaie de détecter toutes les transitions progressives, il en résulte beaucoup de fausses alarmes. Nous pouvons dire que la détection précise des transitions graduelles est encore un problème non résolu. Lienhart [28] a présenté un algorithme de détection d'effet fondu réalisant un taux de détection de 69 % tout en ramenant le taux de fausse alarme à 68 %.
Le but de cette section est de passer en revue quelques travaux existants sur la détection des transitions graduelles. Des aperçus utiles sont également présentés dans certaines recherches [118][35] . Nous pouvons mentionner trois groupes d'approches pour détecter les transitions graduelles qui sont :

1. Approches basées sur les histogrammes.
2. Approches basées sur les contours.
3. Approches basées sur la variance.

2.4.1 Approches basées sur les histogrammes

Une des premières méthodes proposées est la technique de comparaison jumelée proposée par Zhang et al. [32]. Cette méthode compare les différences d'histogrammes avec

deux seuils : un seuil inférieur a été utilisé pour détecter les petites différences qui se produisent pendant la durée de la transition graduelle, tandis qu'un seuil plus haut a été utilisé pour la détection de coupure entre les plans et les transitions graduelles.

2.4.2 Approches basées sur les contours

Pendant une dissolution, les contours des objets disparaissent graduellement tandis que les contours des nouveaux objets deviennent graduellement apparents. Les contours disparaissent graduellement, tandis que pendant une apparition en fondu les contours émergent graduellement. Zabih et al. [120] ont prolongé leur méthode de détection des coupures entre les plans pour détecter les transitions progressives. Ils ont rapporté que le taux de détection des transitions progressives avec cette méthode est bon, mais selon d'autres auteurs [70][66], le taux de fausses alarmes étaient souvent inacceptables. Parmi les raisons des fausses alarmes est le fait que l'algorithme ne compense que pour les mouvements en translation ; un mouvement en zoom conduit directement à une fausse alarme. Lienhart [66] a exploité la perte de contraste des contours pendant une dissolution pour détecter la transition graduelle. Pour ce faire, il a capturé et amplifié la relation entre les contours les plus forts et les plus faibles. Le but de cette méthode a été de résoudre le problème de mouvement de la caméra et des objets rencontré par la méthode de Zabih et al. [120]. Cependant, Lienhart [66] a rapporté que le taux de fausses alarmes reste toujours très haut. Une autre méthode intéressante est la détection des transitions progressives par l'analyse de tranche temporelle [79][80]. La vidéo est représentée comme un volume 3-D qui se compose d'un ensemble de tranches 2-D. Ces tranches ont été alors utilisées pour extraire un indicateur qui peut être utilisé pour capturer la similarité entre les vidéos. Chaque tranche contient les régions de couleur et de texture uniformes et les bordures de ces régions sont utilisées pour détecter la présence de transitions de plan.

2.4.3 Approches basées sur la variance

Une autre méthode pour détecter les transitions progressives est d'analyser le comportement temporel de la variance des intensités de pixels dans chaque image. Ceci a été proposé pour la première fois par Alattar [3]. Ensuite d'autres auteurs [25][69][106] ont proposé des modifications à cette méthode. Alattar [3] a exploité le fait que la courbe de variance dans un effet de dissolution idéale a une forme parabolique. Ainsi, détecter

les effets de dissolution revient à détecter ce modèle dans la série de temps. Même si ces modèles ont une bonne performance, ils sont handicapés par la supposition faite à propos des transitions, car cette dernière ne se généralise pas aux séquences vidéo réelles. Pour résoudre ce problème, Nam et Tewfik [77] ont proposé une technique pour estimer la courbe de transition actuelle par l'utilisation de la technique d'adaptation de courbe polynomiale B-Spline. D'autres auteurs [66][69][25] ont proposé des approches pour la détection des transitions de type fondu.

2.5 Conclusion

Dans ce chapitre, nous avons donné une vue d'ensemble sur les différentes méthodes existantes pour la segmentation temporelle des vidéos. Nous pouvons conclure que la détection des transitions graduelles est plus difficile en raison d'une faible différence entre deux images successives et le nombre inconnu d'images dans la transition [32]. Pour cette raison, nous nous somme intéressée dans les prochains chapitres qu'aux méthodes d'extraction d'images clés qui ne nécessite que l'information que procurent les transitions brusques.

Chapitre 3 :généralités sur les méthodes de cartes de saillances (saliency map)

Chapitre 3

généralités sur les méthodes de cartes de saillances (saliency map)

3.1 Introduction

Dans le chapitre précédant, nous avons donné une vue d'ensemble sur les différentes méthodes existantes pour la segmentation temporelle des vidéos, dans ce chapitre nous nous intéressons aux différentes méthodes de modélisation de l'attention visuelle, à savoir, les méthodes basées sur les cartes de saillances, pour la sélection des informations visuelles spatio-temporelles parmi les plans d'une vidéo qui seront utilisées par la suite pour la sélection des images clés.

3.2 Perception et attention visuelle

La perception visuelle regroupe les mécanismes de mis en œuvre pour la réception et la cognition de stimuli visuels. La partie réceptive se charge de capter et d'organiser les informations visuelles en provenance de l'environnement, alors que la partie cognitive se charge de l'interprétation de ces informations. La réception des stimuli visuels est effectuée par l'œil (la rétine en particulier). La quantité d'informations provenant de ces stimuli visuels étant trop importante pour être traitée dans sa totalité, un mécanisme de sélection des informations est nécessaire : c'est le rôle de l'attention visuelle. Cette sélection s'effectue soit de façon bas niveau, au niveau de la fovéa [111] par exemple , soit de façon plus haut niveau, recherche active d'informations spécifiques par exemple . On

s'intéresse ici principalement aux mécanismes de réception de stimuli visuels ainsi qu'à la focalisation de l'attention visuelle.

3.3 Caractéristiques biologiques de l'œil et modèles psycho-visuels de l'attention

Afin d'être réalistes, les différentes méthodes de focalisation de l'attention visuelle se basent sur des caractéristiques biologiques de l'œil humain ainsi que sur des propriétés psycho visuelles.

3.3.1 Caractéristiques biologiques de l'œil humain

L'œil est l'organe permettant la réception des stimuli visuels. Il est donc important de comprendre son fonctionnement lorsque l'on s'intéresse à la perception visuelle et à la focalisation de l'attention.

Structure globale de l'œil

L'œil est composé d'une partie transparente (cornée, cristallin, humeur aqueuse et corps vitré) permettant au flux lumineux (des photons) d'être acheminé jusqu'à la rétine au fond de l'œil. Le cristallin joue le rôle de lentille convergente afin qu'une image réelle se forme au niveau de cette rétine où sont situés les récepteurs photosensibles.

La rétine est un tissu neuronal qui se charge de transformer un flux lumineux en influx nerveux qui pourra ainsi ensuite être interprété par le cerveau. Pour ce faire elle est composée de plusieurs couches de cellules. La plus profonde de ces couches est constituée de photorécepteurs qui ont la capacité de réagir à un flux lumineux spécifique en envoyant un influx nerveux vers une couche supérieure. La couche suivante (appelée couche granuleuse interne [111]) permet d'une part une transmission directe des informations (grâce aux cellules bipolaires reliant plusieurs photorécepteurs entre eux), et d'autre part une transmission modulée "latéralement" permettant de tenir compte des informations du voisinage (pour mieux s'adapter au contraste ou au contour des objets par exemple). La rétine est plus sensible en son centre (au niveau de la tache orange) et présente à sa surface une légère dépression au niveau de laquelle la perception est plus détaillée (Fi-

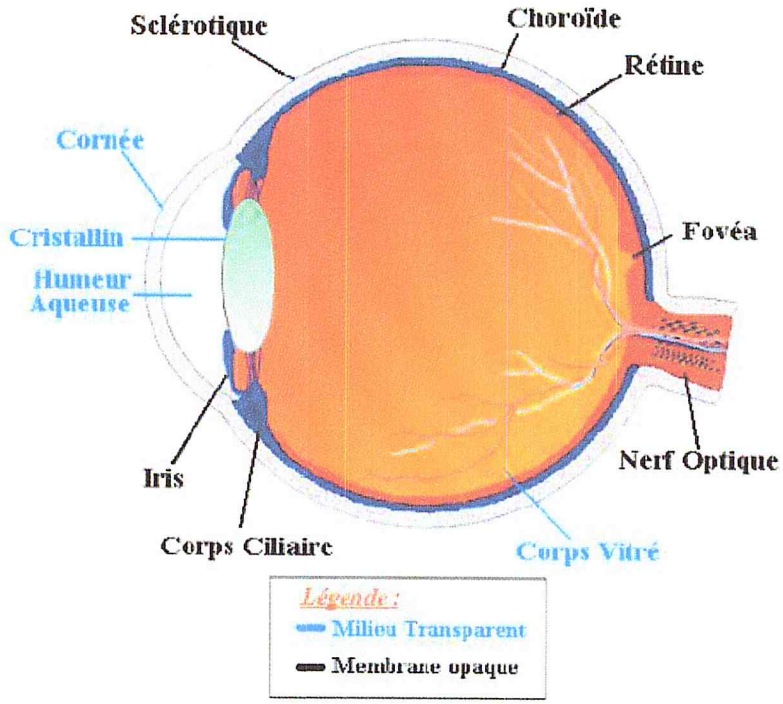


FIGURE 3.1 – Structure globale de l'œil humain.

gure 3.2) : la fovéa, d'où provient la plus grande partie de l'information visuelle transmise au cerveau. La vision périphérique est quant à elle moins précise car la densité de cellules photosensibles est plus importante à proximité de la fovéa. Un phénomène de saccade permet alors de résoudre ce problème de non-uniformité en permettant à l'œil de se fixer (se focaliser) sur les objets les plus intéressants (en amenant les objets en question à être perçus au niveau de la fovéa).

3.3.2 Caractéristiques de la vision naturelle

La biologie ne suffit pas à expliquer l'ensemble des mécanismes de focalisation, d'autres approches cognitives sont nécessaires

Routines visuelles

Des expérimentations dans le domaine de l'attention visuelle, ont permis de déterminer la présence de routines visuelles [36]. Ces routines visuelles sont en fait des schémas d'attention visuelle spécifiques aux différents sujets et liés au but instantané du sujet (c'est à dire ce qu'il doit faire dans les instants qui suivent, la tâche qu'il doit réaliser). Ces schémas d'attention peuvent influer sur différents paramètres de la vision, tels que,

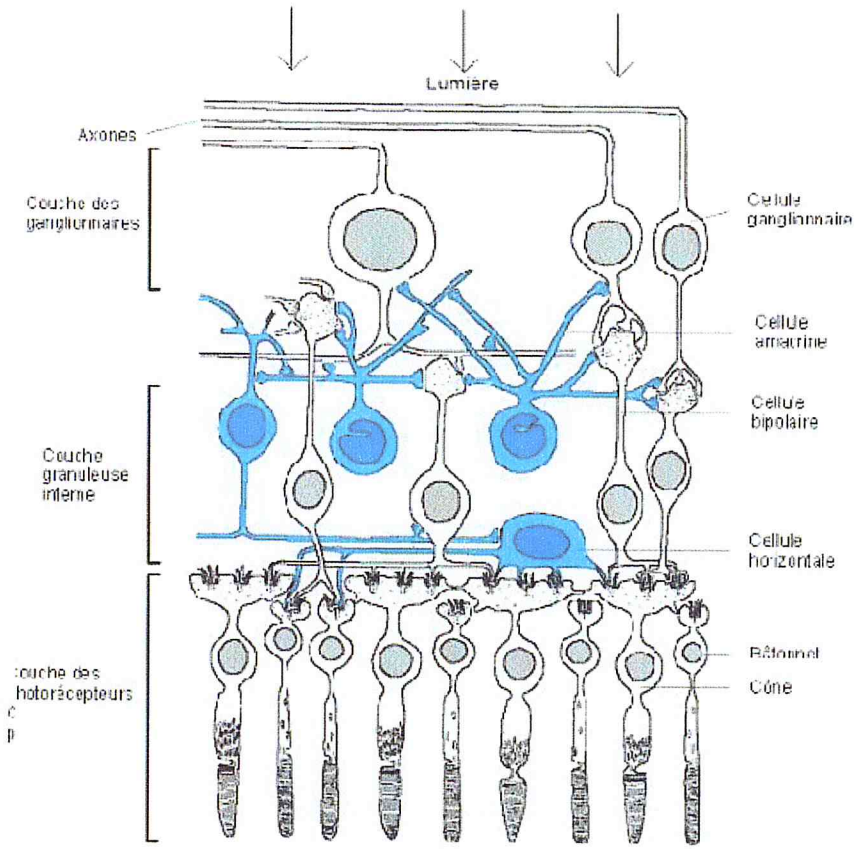


FIGURE 3.2 – Structure de la rétine.

les stratégies d'exploration de l'environnement par le système visuel, la durée de fixation d'un stimulus par exemple. Le déclenchement de ces routines est principalement dû à une recherche active (peu influencée par des stimuli visuels) du sujet. Cette recherche semble s'effectuer suivant un agenda interne spécifique au sujet et dépendant d'une part des tâches qu'il doit effectuer et d'autre part de ses connaissances antérieures (son expérience)[111] .

Sensibilité aux changements

Une des autres caractéristiques de la vision naturelle est la sensibilité aux changements [36][37]. En effet la capacité à détecter des variations entre deux fixations d'une scène est une part importante de la vision car elle permet de détecter des informations inattendues et qui peuvent se révéler importantes relativement au contexte de l'observation (« change blindness ») [93]. Cette sensibilité aux changements dépend de nombreux paramètres.

Sensibilité aux changements

La sensibilité aux changements amène une notion importante dans les mécanismes de vision : la mémoire visuelle. Des théories telles que celle du "change blindness"[101] impliquent une mémoire très limitée d'une fixation sur l'autre (représentation très peu détaillée de la scène sans information particulière sur l'identité des objets par exemple). D'autres expérimentations ont montrés que la mémoire visuelle ne serait pas aussi peu détaillée que le laissait supposer le "change blindness", ainsi on conserve plusieurs informations sur la scène observée entre deux fixations successives. Des paramètres tels que l'identité des objets, leur localisation ou encore la structure spatiale de la scène (afin de prévoir les mouvements à effectuer pour réaliser une tache par exemple) semblent faire partie intégrante de la mémoire visuelle.

3.4 Modèles informatiques

L'attention visuelle humaine désigne le mécanisme de sélection des informations visuelles spatio-temporelles du monde visible. Étant donné que l'environnement visuel contient beaucoup d'informations. Le système visuel humain étant de plus intrinsèquement limité en capacité de traitement, ce dernier s'est adapté par le biais du mécanisme de l'attention visuelle pour réduire la quantité d'information à traiter et pour ne conserver que les informations les plus importantes. En d'autres termes, l'attention visuelle permet d'utiliser de façon optimisée les ressources biologiques ; ainsi, seule une petite partie des informations incidentes est transmise aux aires supérieures du cerveau [23]. William (1890) [104] et Nakayama and Mackeben (1989) [76] avaient émis l'hypothèse d'au moins deux modèles de mécanismes de l'attention, un modèle d'attention ascendant (Bottom-Up), et un modèle d'attention descendant (Top-Down). Ces deux modèles peuvent être décrits comme suit :

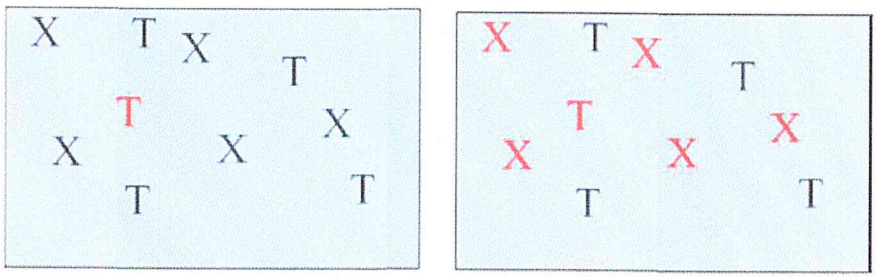
3.5 Les modèles ascendants (bottom-up)

Dans ces modèles, l'analyse des éléments se fait à partir des propriétés de la scène perçue. Ce sont des processus automatiques sélectionnant les informations visuelles selon leur saillance. Ce mécanisme de sélection se fait donc sans aucune connaissance a priori sur la scène. En effet, même en l'absence de tâche à effectuer, le regard se balade lors

d'observation d'une scène et décrit un parcours que ces modèles s'attachent à analyser et à définir ((Figure 3.3)a).

3.6 Les modèles descendants (top-down)

Ce sont des processus contrôlés. Dans certains cas c'est le cerveau lui-même qui envoie directement l'information vers les systèmes sensoriels. Ainsi les modèles d'attention Top-Down sont des modèles principalement dirigés par les connaissances a priori sur la scène ((Figure 3.3)b). En d'autres termes, ces mécanismes sont pilotés par la tâche à effectuer (Taskdependent). La Figure 3.3 donne, pour un même observateur et pour cinq taches différentes à effectuer, les stratégies visuelles associées à la tâche. Ces deux processus sont indispensables pour pouvoir interpréter des scènes en temps réel. Le mécanisme étudié dans ce présent mémoire est le « Bottom-Up ». En effet, afin de réduire la complexité du processus de mise en correspondance des images, et augmenter la robustesse et l'efficacité de ce processus, l'extraction dans une image d'informations pertinentes et robustes, vis-à-vis des éventuels changements comme le point de vue ou les conditions d'acquisition, sans aucune connaissance préalable est souhaitée.



(a) Trouver la lettre rouge. (b) Trouver la lettre « T » rouge.

FIGURE 3.3 – Trouvez le « T » rouge, (a) impossible de ne pas le voir : processus Bottom-Up (cas disjonctif = traitement parallèle) ; (b) Plus difficile qu'en (a) car cela demande processus actif mettant en jeu diverses « stratégies » perceptives : processus Top-Down (cas conjonctif = traitement série) [61].

3.7 Différentes possibilités de simulation de la perception visuelle

Afin de simuler la perception visuelle, deux approches principales se distinguent : l'interrogation directe d'une base de données contenant des informations relatives à l'environnement de l'agent [94][92] [81][56][9] et les méthodes dites de vision synthétique [42][18] [44][46] [112] [86] .

3.7.1 Extraction à partir de bases de données

Les méthodes utilisant un accès direct à une base de données sont relativement rapides à mettre en œuvre dans des cas bien définis (petites scènes par exemple), et peuvent être modulées afin d'être plus réaliste [94]. La perception s'effectue ici par interrogation d'une base de données 3D (la scène) mettant en œuvre des procédures de lancer de rayon, de découpage 3D (clipping[81][56]) et de filtrage sur les propriétés sémantiques des objets de la base[9].

3.7.2 Vision synthétique

La vision synthétique se base sur l'interprétation d'images afin de simuler la perception visuelle. Ainsi dans les méthodes de vision synthétiques, on travaille à partir d'une matrice de pixels, auxquels sont associés des informations (par exemple, une couleur RVB, une profondeur ou encore des informations sémantiques)[111].

3.8 Méthodes associés à la vision

Les différentes méthodes de perception visuelle, outre l'approche utilisée afin d'obtenir des informations sur la scène (extraction à partir de bases de données ou vision synthétique) peuvent se distinguer par la façon dont elles exploitent ces informations. Ainsi certaines méthodes dites :

- bottom-up : des méthodes sont basées uniquement sur des stimuli visuels
- top-down : des méthodes basent principalement sur des informations de haut niveau (buts qui guident une recherche active d'objets particuliers).

- Hybrides : des méthodes combinent les approches bottom-up et top-down afin d'obtenir un résultat plus réaliste et pouvoir faire un tri des informations bas-niveau guidé par des buts.

3.8.1 Approches Bottom-Up

Les approches dites bottom-up se basent principalement sur les modèles biologiques afin de simuler une perception visuelle bas niveau. Elles permettent de simuler une focalisation de l'attention visuelle en tenant compte uniquement des informations reçues au niveau de l'œil (aucun phénomène cognitif) et de la structure de celui-ci. Nous allons présenter ici différentes méthodes utilisant cette approche. Cartes de saillances, les cartes de saillances permettent de regrouper au sein d'une image différentes informations sous la forme d'un scalaire associé à chaque pixel de l'image. Cette valeur est obtenue à partir de différents filtrages appliqués à l'image ou autres paramètres et représente une importance que l'on donne à certaines informations du champ visuel.

Ainsi, à partir d'une image RVB classique (et parfois d'autres images), on extrait différentes informations (différents aspects de l'image) que l'on recombine ensuite permettant ainsi de mettre en valeur certaines zones en fonctions des éléments qui y sont présents . On obtient ainsi une nouvelle image codant la saillance des différents aspects pris en compte et pouvant être utilisée afin de déterminer où l'attention visuelle devrait logiquement se diriger. Courty et al.[18] ont proposé une méthode dont le but est de permettre à un agent autonome de diriger son regard (mouvements de la tête principalement) vers des zones importantes de son environnement et ainsi d'adopter une "attitude" plus réaliste de focalisation spontanée.

Leur méthode utilise un modèle de vision synthétique utilisant une carte de saillances, Ici la focalisation de l'attention s'effectue au moyen de la carte de saillances associée à l'algorithme de WTA [14]. En effet, les zones les plus saillantes sont considérées comme étant des endroits importants vers lesquels l'attention doit se focaliser (où le regard se dirige) et l'algorithme WTA permet alors de retrouver cette zone. Ce principe de focalisation de l'attention basé sur les cartes de saillances est efficace dans le premier instant de fixation d'une scène mais devient relativement limité lors de l'analyse plus approfondie de la scène (focalisation spontanée). D'autres méthodes utilisent le principe de carte de saillances associé à un algorithme WTA. Ces méthodes diffèrent principalement dans la

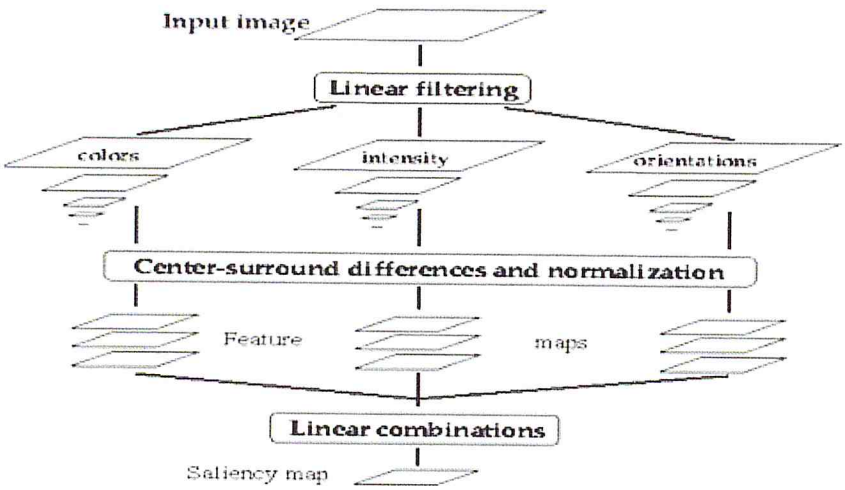


FIGURE 3.4 – Un modèle de carte de saillances [46].

nature des caractéristiques qui sont extraits de l'image de départ ainsi que dans la façon dont ils sont combinés au sein de la carte de saillances. Ainsi, dans [46][44][43] des caractéristiques de couleur, d'intensité ou encore d'orientation sont pris en compte. Différentes façons de combiner ces caractéristiques influent ensuite sur le réalisme du modèle de carte de saillances utilisé, ainsi Itti et Koch [45] comparent différentes méthodes de combinaison (linéaire simple, en affectant différents poids, ...) et proposent finalement une méthode itérative permettant de prendre en compte certaines caractéristiques locales (qui s'annulent mutuellement lorsque l'on utilise d'autres méthodes de combinaison). Sun et Fisher [103] proposent quant à eux un modèle prenant en compte les relations spatiales entre et au sein même d'objets. La méthode proposée par Terzopoulos et al.[112] a pour objectif de proposer un système de vision généraliste , Ce système de vision utilise tout d'abord uniquement des images couleur (RVB) afin de réaliser l'acquisition des informations visuelles. Au niveau de chaque œil, quatre caméras coaxiales permettent de simuler la projection du monde environnent sur la rétine. Ces quatre caméras travaillent en projection perspective et à des résolutions différentes ce qui permet donc d'une part, de tenir compte de l'occlusion entre les objets, et d'autre part de simuler la répartition non uniforme des cellules photosensibles (vision détaillée au centre et plus "floue" à la périphérie). les modèles connus (représentés par des histogrammes de couleurs [105]) afin de déterminer ce qu'il doit fixer dans son environnement. Pour ce faire, on utilise avant tout une comparaison entre les histogrammes de couleurs servant de modèles et

l'image perçue, ainsi on peut déterminer la présence (ou non) du modèle au sein de l'image. Ensuite, on effectue une projection inverse (en affectant un poids élevé aux pixels de l'image correspondant le plus à l'histogramme recherché), qui permet de déterminer où se situe la cible dans l'image. Une fois l'objet de la fixation déterminé, les yeux effectuent un mouvement de saccade [111] (en incrémentant les variables contrôlant l'orientation des yeux) afin d'amener l'image de cet objet au niveau de la fovéa. Ensuite une stabilisation de la vision est effectuée (nécessaire du fait de l'ondulation du corps du poisson virtuel) grâce à une comparaison entre les schémas d'intensité des instants t et $t1$. La méthode proposée par Peters et Sullivan [86] se base sur une combinaison de techniques de rendu, d'accès à une base de données et de cartes de saillances afin d'obtenir un modèle d'attention spontanée pouvant être utilisé seul (attention spontanée d'un agent virtuel pour diriger son regard comme dans [18]) ou par exemple pour interrompre un processus d'attention de plus haut niveau (guidé par des buts) afin de prendre en compte des événements inattendus. Le modèle se décompose en plusieurs étapes. Tout d'abord, on effectue trois rendus différents du point de vue de l'agent : un rendu complet (rendu avec éclairage et texture) qui est utilisé pour la construction de la carte de saillances et deux autres rendus à des résolutions plus faibles et sans informations d'éclairage ou de texture qui seront utilisés dans le cadre d'une forme de mémoire (comme dans [56]). Ces deux derniers types de rendu permettent également d'effectuer une approximation de l'acuité visuelle . Une fois ces différents rendus effectués, le module d'attention visuelle utilise l'image résultant du rendu complet pour construire une carte de saillances décrivant l'importance des stimuli visuels perceptibles par l'agent. Les auteurs de cette méthode utilisent un modèle de carte déjà existant [46][44][43] prenant en compte des informations d'orientation, de couleur et d'intensité. Les différentes caractéristiques prisent en compte sont ensuite combinés grâce à un opérateur de normalisation non-linéaire [45] permettant de donner plus d'importance aux caractéristiques moins nombreuses (et donc a priori plus importantes).

3.8.2 Approches Top-Down

A l'inverse des approches vues précédemment, les approches top-down [31], essayent de tenir compte des intentions (buts immédiats) et font donc de ce fait intervenir des concepts de plus haut niveau que les stimuli visuels. Ces approches permettent de tenir compte des modèles psycho-visuels et amènent donc également leur part de réalisme.

La méthode proposée par Bordeux et al.[9] n'utilise pas la vision synthétique mais l'accès direct aux informations de différentes bases de données. Les auteurs proposent d'appliquer à la base de données contenant la scène une suite de filtres. Chacun de ces filtres ne laisse passer qu'un nombre réduit d'objets. Il est ainsi possible par exemple de ne sélectionner que les objets correspondants à des voitures rouges se dirigeant vers un point précis. Dans un tel exemple, les filtres utilisés seraient "voiture", "rouge", "orientation",... etc Pour ce faire, leurs agents sont "équipés" d'un ensemble de flux de perception, chaque flux étant lui-même composé de différents filtres. Ainsi différents filtres peuvent être envisagés tels qu'un filtre de perception de la distance ou un filtre limitant la perception au champ de vision de l'agent. Ces filtres effectuent un tri sur un ensemble de données qui leur sont fournies suivant certaines propriétés (par exemple, tous les objets comportant une face carrée). Ces filtres peuvent ensuite être associés au sein d'un pipeline correspondant à un certain type de perception.

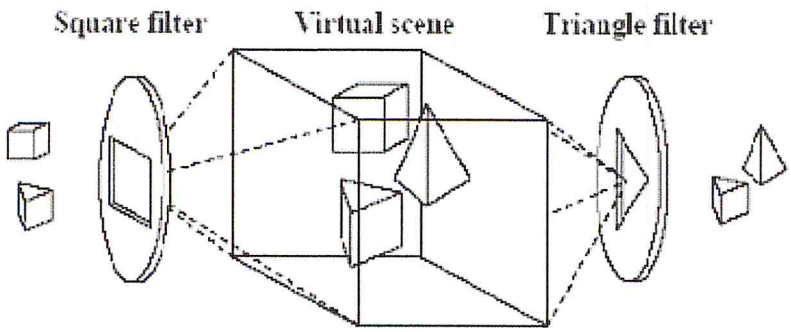


FIGURE 3.5 – Un exemple de filtres.

Ici la perception est essentiellement guidée par les buts des agents (les différents types de filtres appliqués dépendent de ces buts ainsi que du type de perception). En termes de temps de calcul, les différents tris au niveau des bases de données sont aussi longs que les filtrages appliqués par exemple dans le cas de cartes de saillances complexes [46]. La focalisation de l'attention est ici effectuée directement au niveau des bases de données (tri par rapport au champ de vision par exemple), celle-ci reste donc assez peu réaliste (par rapport aux conditions d'éclairage par exemple). La méthode que proposent Kuffner et Latombe [56] repose sur une combinaison de vision synthétique et d'accès direct à une base de données. La partie vision synthétique sert ici à déterminer la visibilité des objets (comme dans [86][81]). Pour ce faire un rendu simple (sans éclairage ni texture)

est effectué et l'image résultante de ce rendu est analysée. Une couleur unique étant attribuée à chaque objet, cette analyse permet de retrouver les objets vus par un agent. Une fois ces objets déterminés, l'agent analyse certaines propriétés de ces objets (propriétés dépendant du but de cet agent, ce à quoi il doit s'intéresser) et choisi alors de mémoriser des informations relatives aux objets qui sont importants pour lui. Des informations telles que l'identifiant de l'objet en question, certaines de ses propriétés, sa vitesse, sa position et son orientation, ou encore le moment de l'observation par exemple. Ainsi, une forme de perception visuelle (rendu simple) et de focalisation de cette perception (choix parmi les objets visibles ceux qui sont intéressants) sont mis en place. Cette focalisation est intéressante parce qu'elle permet de faire directement un lien entre des notions de haut niveau et des stimuli visuels. Malgré tout, elle ne donne pas d'importance particulière aux zones au sens de l'importance de certains stimuli forts dans la réalité (stimuli qui ne sont pas forcement importants au regard de la tache à effectuer). Ceci limite fortement les possibilités de focalisation spontanée de l'attention visuelle. La méthode proposée par Baluja et Pomerleau [88] permet de piloter automatiquement un véhicule à partir d'une séquence d'images représentant la route devant lui. Le véhicule est contrôlé par un réseau de neurones dont la couche d'entrée est connectée à une caméra et la couche de sortie à un dispositif contrôlant la direction du véhicule. Cette approche simple n'est pas envisageable dans un environnement réel car le réseau de neurones est sensible aux bruits (piétons, arbres, bâtiments,...). Afin de réduire ce bruit, des couches internes du réseau de neurones sont utilisées afin de coder une carte de saillances. Celle-ci permet de se focaliser sur les zones importantes dans la séquence d'images acquise par la caméra du véhicule.

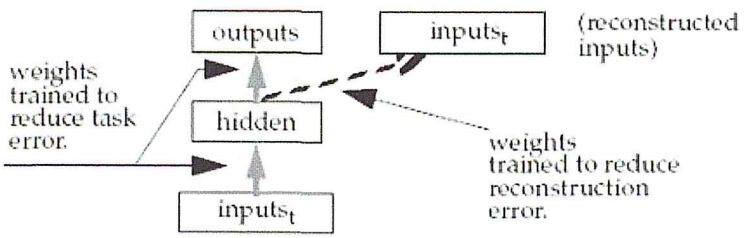


FIGURE 3.6 – Le réseau de neurones et la carte de saillances

Il s'agit d'une focalisation guidée par les buts car le réseau de neurones est entrainé à une tache : la focalisation ne peut se faire que sur des motifs appris. La carte de saillances trouve ici une nouvelle application et permet d'aider au filtrage des informations en réduisant la sensibilité au bruit du modèle basé sur le réseau de neurones.

3.8.3 Approches hybrides

Des méthodes cherchent à combiner des approches ascendantes et descendantes de façon à pouvoir réagir aux événements inattendus tout en se focalisant sur des buts définis à priori. Le modèle proposé par Yaoru Sun et Robert Fisher [103] adapte les précédents modèles à base de carte de saillances afin de prendre en compte une notion de groupes (un groupe pouvant être une zone de l'image analysée, un groupe d'objets ou encore des points par exemple). Cette partie bottom-up du modèle se voit ensuite complétée par un module permettant de guider l'attention en fonctions de buts prédéfinis. Le modèle proposé par Olivia et al.[82] utilise une approche hybride en modulant un modèle de carte de saillances en fonction de buts. Pour ce faire, une notion de probabilité de présence est associée la carte de saillances afin de privilégier des zones dans lesquels des objets correspondants aux buts sont le plus susceptible d'apparaitre. Ici, on extrait d'abord différents aspects de l'image en décomposant chaque information de couleur à l'aide d'un filtre (pyramide orientable en utilisant quatre niveaux d'échelle et quatre orientations). A chaque point est ensuite associé un vecteur de caractéristiques, composé de 48 valeurs $((3couleurs)(4echelles)(4orientations))$. Puis, on définit la saillance comme étant la probabilité de trouver un ensemble de caractéristiques locales dans l'image. Ainsi, la saillance d'un point est plus importante lorsque des informations présentes en ce point sont peu "attendues" (on une faible probabilité d'apparaitre). Cette probabilité est ensuite modulée par une fonction gaussienne. Le modèle proposé par Navalpakkam et Itti[78] a pour objectif d'extraire d'une scène toutes les informations qui paraissent importante suivant un but déterminé. Pour ce faire, les auteurs utilisent une combinaison entre une approche bottomup basée sur une carte de saillances et une approche top-down avec la création d'une carte de taches. La carte de saillances est en fait ici basée sur le modèle utilisé dans plusieurs autres méthodes ([46][42]). La carte de taches est quant à elle, une image permettant, une fois combinée à la carte de saillances, de privilégier ou d'inhiber certaines zones en fonction d'un objectif particulier au sein d'une carte de guidage de l'attention (par exemple si l'on demande de retrouver les voitures dans une scène, alors les zones correspondant à la route seront privilégiées alors que les immeubles seront plutôt inhibés). Cette méthode est donc basée sur quatre composants principaux, le Visual brain [93] qui se charge de maintenir les différentes cartes (saillance, tâche et guidage de l'attention), la mémoire de travail (working memory) [6] qui s'occupe de la création et du maintien

à jour d'un graphe de taches (utilisé dans la création de la carte de taches), la mémoire à long terme (ontology) [4] et enfin l'agent qui se charge de faire communiquer les différentes entités vues précédemment. Tout d'abord, le visual brain reçoit des informations en provenance d'un moteur de rendu (ou autre source d'image, dans le cas présent cette source est une vidéo). Puis, il extrait différentes caractéristiques de cette image afin de créer une carte de saillances (du même type que les cartes utilisée dans [46][44]).

3.9 Conclusion

Dans ce chapitre, nous avons donné une vue d'ensemble sur les différentes méthodes existantes pour la modélisation de l'attention visuelle dont le but est l'extraction des cartes de saillances, nous présenterons dans le prochain chapitre la méthode retenue pour l'extraction des images clés à partir d'une vidéo.

Chapitre 4 :la conception et la réalisation

Chapitre 4

Méthodologie et Résultats

4.1 Introduction

Dans les chapitres précédents nous avons donné des notions de bases sur la vidéo et ses caractéristiques, nous avons présenté également quelques méthodes de segmentation de vidéo en plan ainsi que quelques méthodes d'extraction de la carte de saillance à partir d'une image, nous allons voir dans ce chapitre comment nous avons combiné ces méthodes afin d'extraire les images clés à partir d'une vidéo.

4.2 Méthodologie

Nous avons implémenté deux méthodes pour l'extraction des images clés, La première méthode de résumé vidéo s'appuie sur des caractéristiques de bas niveau, à savoir, l'histogramme d'ordre 0, et la moyenne, nous étudierons plus particulièrement la manière de les combiner pour créer le résumé vidéo [32]. L'autre méthode repose sur des caractéristiques de plus haut niveau comme les histogrammes d'ordres supérieurs à 0 et les cartes de saillances [7].

4.2.1 Première approche pour l'extraction des images clés

Cette méthode est proposée par Poonam S.Jadhav[47], elle est basée sur la comparaison des histogrammes d'ordre 0 pour l'extraction des images clé de la vidéo, son principale avantage est sa faible complexité algorithmique. Son schéma synoptique est représenté dans la Figure 4.1

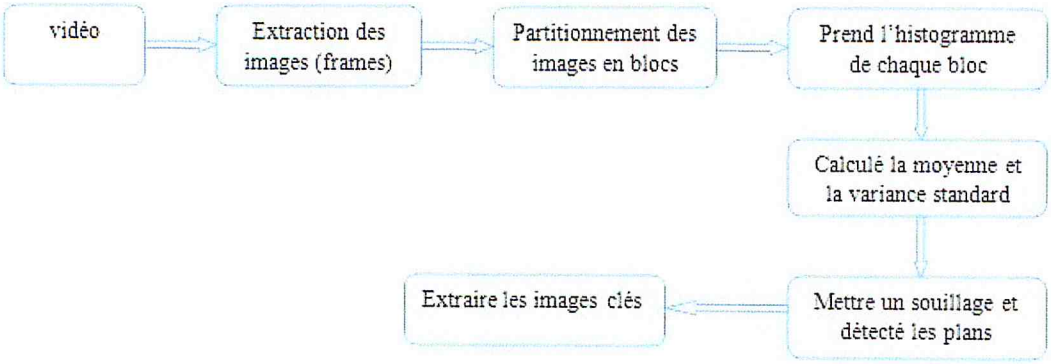


FIGURE 4.1 – schéma synoptique de la méthode de Poonam S.Jadhav[47].

Segmentation de la vidéo en plan

Cette étape a pour objectif d'extraire les différents plans de la vidéo. Cette segmentation se fait en plusieurs étapes :

- **Partitionner les images en blocs :** Dans un premier temps, chaque image de la vidéo est partitionnée en $m \times n$ blocs, telle que $B(i,j,k)$ représente le bloc (i,j) dans la $k^{ème}$ image comme illustré dans la figure 4.2

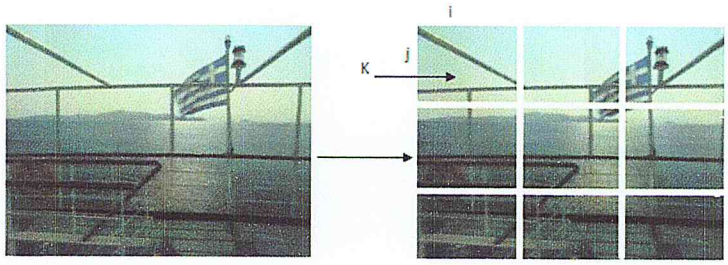


FIGURE 4.2 – Partitionnement d'une image en blocs .

- **L'histogramme de différence X^2 entre les blocs :** Pour chaque couple d'images successives de la vidéo, l'histogramme de différence X^2 [83] est calculé entre les blocs correspondants. soient $H(i,j,k)$ et $H(i,j,k+1)$ les histogrammes du bloc (i,j) pour la $k^{éme}$ et $(k+1)^{éme}$ image, respectivement. La différence entre les blocs est calculée par l'équation suivante :

$$D_B(K,K+1,i,j) = \sum_{l=0}^{L-1} \frac{[H(i,j,K) - H(i,j,K+1)]^2}{H(i,j,K)} \tag{4.1}$$

Où L est le nombre de niveau de gris dans une image.

- **L'histogramme de différence X^2 entre deux images :** L'histogramme de différence X^2 entre deux images consécutives d'une vidéo est calculé comme suit :

$$D(K, K+1) = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} D_B(K, K+1, i, j) \tag{4.2}$$

Où w_{ij} représente le poids du bloc (i, j) qui est choisi aléatoirement La différence $D(k, k+1)$ est utilisé pour calculer la variance moyenne et l'écart type sur toute la séquence vidéo [22] comme suit :

$$MD = \frac{\sum_{k=1}^{Fv-1} D(K, K+1)}{Fv - 1} \tag{4.3}$$

$$STD = \sqrt{\frac{\sum_{k=1}^{Fv-1} (D(K, K+1) - MD)^2}{Fv - 1}} \tag{4.4}$$

- **Extraction des transitions brusques :** Les transitions brusques de la vidéo sont extraites par un seuillage automatique, le seuil T est calculé comme suit :

$$T = MD + a \times STD \tag{4.5}$$

Avec a un nombre aléatoire entier Si la condition $D(i, i+1) \geq T$ est vérifiée alors l'image i est considérée comme la fin du plan précédent, et l'image $(i+1)$ est considérée comme le début du nouveau plan.

Sélection des images clés

La détection des images clés se fait par une comparaison entre les images du même plan, pour chaque plan une seule image sera considérée comme image clé. La sélection de l'image clé dans le plan se fait de telle sorte que l'image retenue puisse vérifier les deux conditions suivantes :

1. le maximum de la moyenne dans le plan.
2. le maximum de la variance dans le plan.

$T(i)$ =image clé si $T(i) = max(std)$ et $max(MD)$
i c'est l'indice de limage T telle que $i \in (1 \ldots n)$.

Résultats et interprétations

Afin d'évaluer les performances de la méthode proposée par Poonam S.Jadhav[47], nous l'avons implémenté puis nous l'avons testé sur une sélection de vidéos de la base de données Vidéo SUMMarization (VSUMM) [1]. VSUMM est une base de données crée par des chercheurs, elle contient différent types de vidéos en format MPEG leur durée varie de 1 à 4 minutes, chaque vidéo dispose de cinq résumés vidéo crée par cinq utilisateurs différents (ces résumé seront utilisés pour les comparaisons des méthodes). Les utilisateurs, sont des personnes qui ont traité manuellement les vidéos de VSUMM pour obtenir des résumés, ce qui signifie que chaque vidéo dispose de 5 résumés vidéo crée par 5 utilisateurs différents. Nous avons utilisé trois mesures distinctes pour évaluer cette technique qui sont : le rappel, la précision et F-mesure. Les résultats obtenus sont présentés dans le tableau 4.1.

4.2.2 Le rappel (Recall)

Le rappel est défini par le nombre d'images clés pertinentes retrouvées au regard du nombre d'images clés de la base de données VSUMM. Il Mesure la capacité de la méthode à donner toutes les solutions pertinentes. Il se calcule comme suit :

$$Recall = \frac{imagescorrectes}{imagescorrectes + imagesoubliées} \tag{4.6}$$

4.2.3 La précision

La précision est le nombre d'images clés pertinentes retrouvées rapportée au nombre total d'images clés VSUMM. Elle permet de mesurer la capacité de la méthode à refuser les solutions non-pertinentes. Elle se calcule comme suit :

$$Précision = \frac{imagescorrectes}{imagescorrectes + imagesFaussesdétections} \tag{4.7}$$

4.2.4 F-mesure

C'est la Moyenne harmonique de la précision et du rappel. Elle mesure la capacité de la méthode à donner toutes les solutions pertinentes et à refuser les autres. Elle se calcule comme suit :

1. https ://site .google.com/site/vsummsite/

$$F - mesure = \frac{2Précision \times Recall}{Précision + Recall} \tag{4.8}$$

Elle est également connue sous le nom de mesure , car la précision et le rappel sont pondérés de façon égale et sa valeur varie entre 0 et 1, plus sa valeur se rapproche de 1 et mieux. Le tableau suivant récapitule les mesures obtenues pour différentes vidéos VSUMM :

vidéo	seuil	plans	images clé	Précision	Recall	F-mesure
v22	0.49	1769	292	0	0	0
v47	0.65	463 696 998	193 523 697	1	0.6	0.75
v48	0.3	1237 1644 1764 1979 2447 3166	677 1238 1761 1866 1980 2624	0.33	0.29	0.3
v50	0.5	703 1739 2294 3824	52 1292 1989 2764	0.5	0.25	0.33
v56	0.75	1254 1352	717 1341	1	0.13	0.23
v57	0.7	1254 1352	717 1341	1	0,40	0,57
v62	0.87	655	631	1	0.25	0.4
v66	0.23	416 770 1305 1696	410 417 1214 1306	1	0.5	0.66
v69	0.5	1156 1489 2278 2665 2960	157 1173 1925 2279 2960	0.8	0.44	0.56
v70	0.43	17 554 888 1128 1300	17 222 888 889 1179	0.33	0.4	0.36

TABLE 4.1 – mesures d'évaluation de la méthode Poonam S.Jadhav[47] sur les vidéo VSUMM.

La figure 4.3 représente les valeurs de F-mesures obtenues.

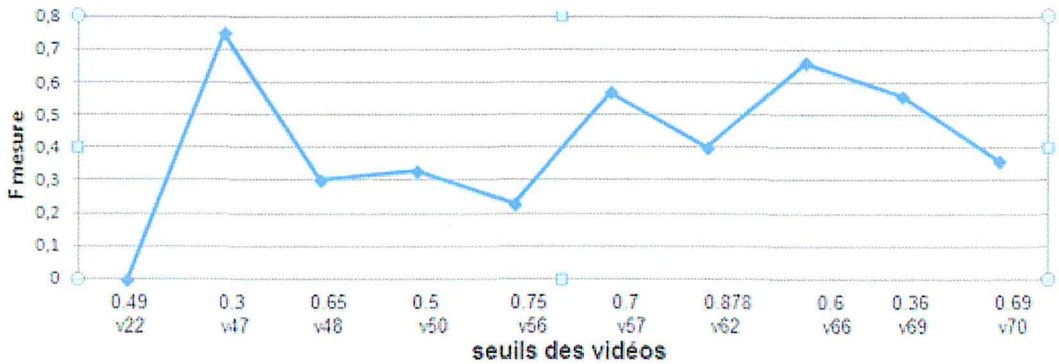


FIGURE 4.3 – F-mesures de la methode de Poonam S.Jadhav[47] sur les vidéos VSUMM .

Les résultats obtenus dans le tableau 4.1 et la figure 4.3 montrent que la valeur maximale de F-mésure est égale à 0.75 et sa valeur moyenne est de 0.5918. cette moyenne est acceptable mais montre que beaucoup d'images clés sont ratées pour les vidéo VSUMM

utilisées. Comme nous ne pouvons pas nous contenter de ce taux moyen dans notre projet de fin d'étude, nous avons essayer de l'augmenter en proposons une nouvelle approche pour l'extraction des images clés.

Deuxième approche pour l'extraction des images clés (approche proposée)

Cette méthode se divise en deux parties, la première partie compare les histogrammes spatiaux pour extraire les transitions brusques. La deuxième partie utilise les cartes de saillances afin de sélectionner les images clés à partir de ces transitions, la Figure 4.4 illustre le schéma synoptique de la l'approche que nous avons proposée pour l'extraction des images clés.

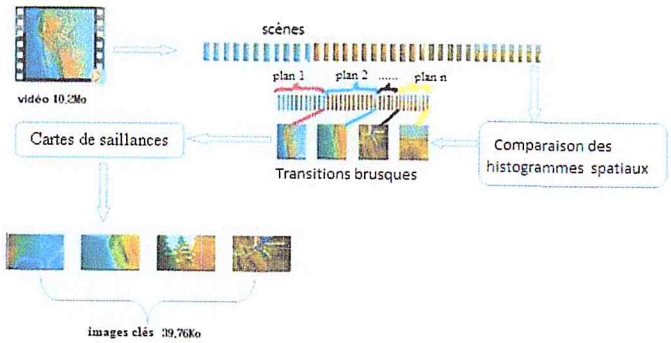


FIGURE 4.4 – schéma synoptique de l'approche proposée.

segmentation de la vidéo en plans

Nous avons choisi d'utiliser la méthode proposée par C.O Conaire, et al [16], pour extraire les plans d'une vidéo, elle se base sur l'utilisation des histogrammes d'ordres supérieures à 0 communément appelés spatiograms [8] pour l'extraction des transitions brusques, ces derniers sont définis comme étant une généralisation d'histogramme qui permet de conserver les informations spatiales telle que la moyenne et la covariance des positions des pixels. La Figure 4.5 illustre les différentes étapes de cette méthode.

- **Les étapes de cette méthode** : Pour détecter les différents plans (les transitions brusques) à partir de la vidéo il est nécessaire de suivre les étapes suivantes :

  1. **Comparaison entre deux Spatiograms :** La comparaison entre deux spatiograms pour chaque couple d'images successive de la scène, se fait en calculant

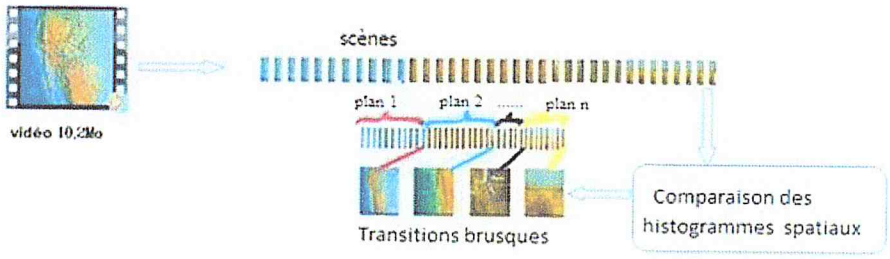


FIGURE 4.5 – Méthodologie de détection des transitions brusques.

l'histogramme normalisé de chaque image cible de N pixels. L'histogramme représente le nombre de fois d'appariation d'un pixel dans une image donc c'est un histogramme spatial d'ordre 0 (cf. Figure 4.6) [126].

Soient $x \in X$ (nombre total de pixel) et $v \in V$ (le nombre total de bins de l'histogramme), $nb(v)$ représente le nombre de pixel x telle que $f(x) = v$. On peut décrire l'histogramme d'ordre 0 nb par d'autre façon comme suit :

$gf(x, y) = 1 \quad si \quad f(x) = v \quad 0 \quad sinon$

La formule suivante définie l'histogramme :

$$nb(v) = \sum_{x \in X} gF(x, v) \tag{4.9}$$

qui est illustré dans la Figure 4.6.

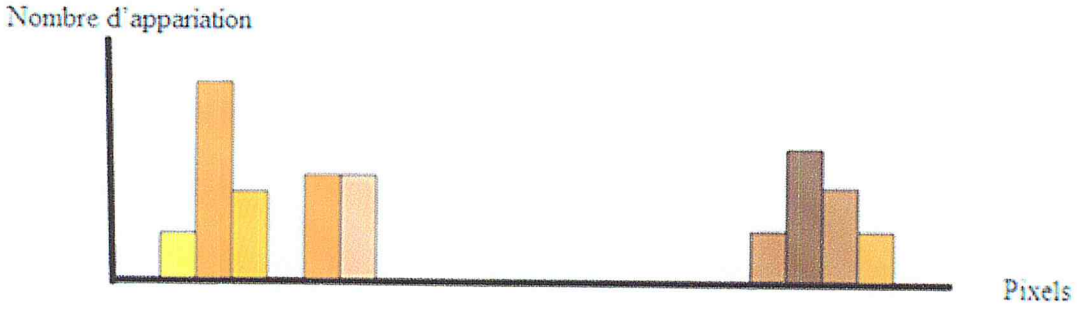


FIGURE 4.6 – Histogramme (Spatiogram d'ordre 0)

Le spatiogram d'ordre 0 ne prend pas en considération les informations spatiales telles que la moyenne et la covariance des pixels, c'est pour cette raison qu'il est impératif de passer à un ordre supérieur qui prend comme paramètre la moyenne et la covariance des positions spatiale des pixels pour chaque histogramme d'ordre 0. Les formules suivantes calculent la moyenne et la covariance :

$$\mu b = \frac{1}{v} * \sum_{i=1}^{x} x_i * gF(x, v) \tag{4.10}$$

$$\sum b = \frac{1}{v} * \sum_{i=1}^{x} (x_i - \mu b) * (x_i - \mu b)^T \tag{4.11}$$

Où $xi = [x, y]^T$ est la position spatiale du pixel i, Pour comparer les différentes tailles des régions, il est nécessaire de mettre les coordonnées spatiales dans une même échelle. Afin de comparer deux « spatiogram » $s = hf, \mu, \sum$ et $s' = hf', \mu', \sum'$, qui prennent l'histogramme, la covariance et la moyenne comme paramètres pour chaque V bins, la formule suivante est utilisée pour la comparaison :

$$\sum_{v=1}^{V} \Phi b \sqrt{hf, hf'} \tag{4.12}$$

$$\Phi = \mu \exp\{-\frac{1}{2}(\mu B - \mu B')^T \sum_{v}^{-1} (\mu B - \mu B')\} \tag{4.13}$$

avec Φ est une mesure spatiale de similarité [84] et μ qui représente le terme de normalisation « Gaussians » Le point faible de cette mesure est qu'il ne tolère pas les petites changements des caractéristiques spatiales telles que les positions des pixels et la taille des bins. Pour résoudre ce problème, une nouvelle mesure de similarité a été proposée, elle consiste à comparer deux spatiograms avec l'utilisation du coefficient de « Bhattacharyya » [15]. C'est une mesure absolue de similarité utilisée pour comparer la ressemblance entre deux histogrammes, mais dans notre cas nous avons affaire à des « spatiograms » donc il faut convertir ces derniers en des histogrammes en ajoutant une dimension d'espace pour la moyenne μ et la covariance $\sum$ des pixels telle qu'illustré dans la Figure 4.7, et la formule de ce nouveau histogramme est donnée comme suit :

$$nb, k = \frac{nb\Phi(k\Delta\omega)\Delta\omega}{\sum_{i=-\infty}^{+\infty} \Phi(i\Delta\omega)\Delta\omega} \tag{4.14}$$

avec k un nombre entier allant de $-\infty$ à$+\infty$.

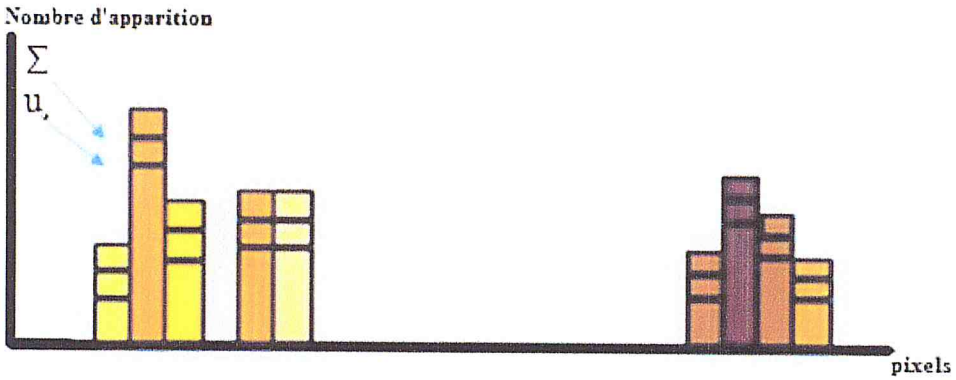


FIGURE 4.7 – Histogramme spatial d'ordre supérieur à 0.

A ce niveau, la comparaison entre deux « spatiogram » se fait en utilisant le coefficient de « Bhattacharyya » qui se calcule comme suit :

$$p(n,n') = \sum_{v=1}^{V} \sum_{k=-\infty}^{+\infty} \sqrt{nb, knb', k}$$

$$= \sum_{v=1}^{V} \sum_{k=-\infty}^{+\infty} \sqrt{\left(\frac{nb\Phi b(k\Delta\omega)\Delta\omega}{\sum_{k=-\infty}^{+\infty} \Phi b(i\Delta\omega)\Delta\omega}\right)} * \sqrt{\left(\frac{nb'\Phi b'(k\Delta\omega)\Delta\omega}{\sum_{k=-\infty}^{+\infty} \Phi b'(i\Delta\omega)\Delta\omega}\right)} \quad (4.15)$$

Où $\Delta\omega$ représente la taille spatiale de chaque bin, et Φb est la normalisation Gaussienne qui se calcule comme suit :

$$\Phi = \mu \exp\{-\frac{1}{2}(\mu B - \mu B')^T \sum_{v}^{-1}(\mu B - \mu B')\} \quad (4.16)$$

$\Delta\omega \longrightarrow 0$.

$$\sum_{i=-\infty}^{+\infty} \Phi(i\Delta\omega)\Delta\omega = \int_{-\infty}^{+\infty} \Phi(x)dx = 1 \quad (4.17)$$

Donc :$p(n,n\prime) = \sum_1 v\sqrt{nb, nb'} \int_{-\infty}^{+\infty} \Phi(x) Phi'(x)dx$

On a $\sqrt{N(x,a,A)} = qN(x,a,2A)$

avec $q = 2(2\pi)^{\frac{m}{4}}|A|^{\frac{1}{4}}$pour m dimension, et $N(x,a,A)N(x,b,B) = Zn(x,c,C)$

avec $Z = N(a,b,A+B)$

$$p(n,n\prime) = \sum_{1}^{v} \sqrt{nb, nb'} \int_{-\infty}^{+\infty} Zb\Phi' v(x)dx \quad (4.18)$$

$$p(n,n\prime) = \sum_{1}^{v} \sqrt{nb, nb'}[qb\ Qb\ x(\mu b, \mu b', 2, (\sum b + \sum b'))] \quad (4.19)$$

$$qb = 2\sqrt{2\pi}|\sum b + \sum b'|^{\frac{1}{4}} \tag{4.20}$$

$$Qb = 2\sqrt{2\pi}|((\sum b)^{-1} + (\sum b')^{-1})^{-1}|^{\frac{1}{4}} \tag{4.21}$$

$$p = \sum_{1}^{v} \sqrt{nb\ nb'}[qb\ Qb\ x(8\pi|\sum b + \sum b'|^{\frac{1}{4}}\mu b, \mu b', 2(\sum b + \sum b'))] \tag{4.22}$$

Où p représente un vecteur de similarité entre les images qui varie entre 0 et 1.

2. **Détection des plans** : Après l'utilisation du coefficient de Bhattacharyya nous avons obtenu un vecteur de similarité qui varie entre 0 et 1, 0 signifie qu'il n'y a pas de similarité et 1 signifie que la similarité est totale. Pour extraire les transitons brusques il suffit d'appliquer un seuillage au vecteur de similarité, le seuil est choisi de telle sorte qu'il puisse tolérer un certain degré de similarité entre les images de la vidéo (cf. Figure 4.8).

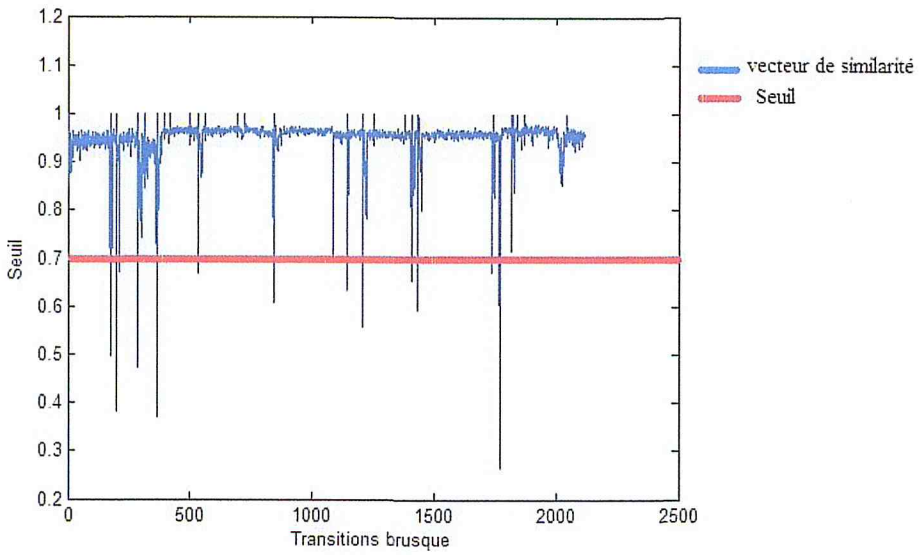


FIGURE 4.8 – Vecteur de similarité .

Une fois que la vidéo est segmentée en plan, nous allons suivre la méthodologie décrite dans la Figure 4.9 afin de sélectionner une image clé de chaque plan.

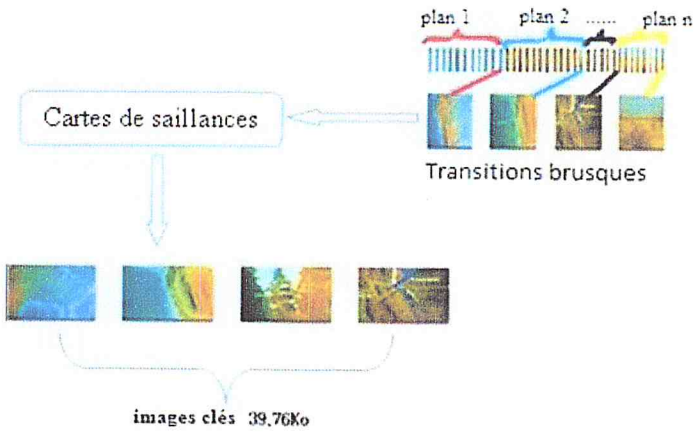


FIGURE 4.9 – Méthodologie de sélection des images clés.

Carte de saillance « Bottom-up »

Nous avons choisi d'utiliser la méthode proposée par Christopher Kanan et Garrison Cottrell[50], elle a pour objectif d'extraire les cartes de saillances des images de la vidéo, elle est basée sur la stratégie de Bottom-up pour guider l'intention vers les objets intéressants dans une image comme illustré dans la Figure 4.10.

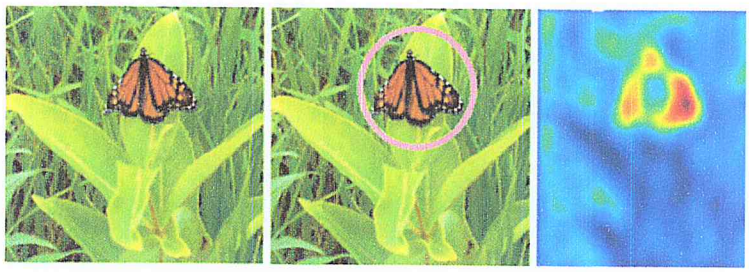


FIGURE 4.10 – de gauche à droite : image initiale, objet d'intérêt, carte de saillance correspondante .

Cette méthode nécessite plusieurs étapes :

- Prétraitement d'images.
- Convertir les images dans espace LMS.
- La normalisation des images LMS.
- Le filtrage Gaussien sur les images LMS.
- Effectué l'analyse en composantes indépendantes (ICA)
- Extraction des cartes de saillance.
- **Prétraitement d'images :** Avant de commencer le prétraitement, il est nécessaire de réduire la taille de l'image a une taille standard. Cela se fait en modifiant la taille

de chaque image telle que sa plus petite dimension ne soit pas inférieure à 128×128.

- **Convertir les images en espace LMS :** Les images doivent être converties de l'espace couleur RGB vers l'espace couleur LMS, ce dernier est conçu pour être similaire aux réponses de l'œil humain [84], la conversation se fait par les équations suivantes :

  $L(\lambda) = 0.214808r(\lambda) + 0.751035g(\lambda) + 0.04515b(\lambda)$

  $M(\lambda) = 0.0222882r(\lambda) + 0.940534g(\lambda) + 0.076827b(\lambda)$

  $S(\lambda) = 0.0000000r(\lambda) + 0.016500g(\lambda) + 0.999989b(\lambda)$

  Avec r, g, b les intensités des couleurs rouge, verte et bleue respectivement, et $\lambda r, \lambda g, \lambda b$ sont des longueurs d'onde de trois couleurs monochromatiques (lumière couleur)r, g, b où $380 \leq \lambda \leq 780nm$ comme illustré dans la Figure 4.11.

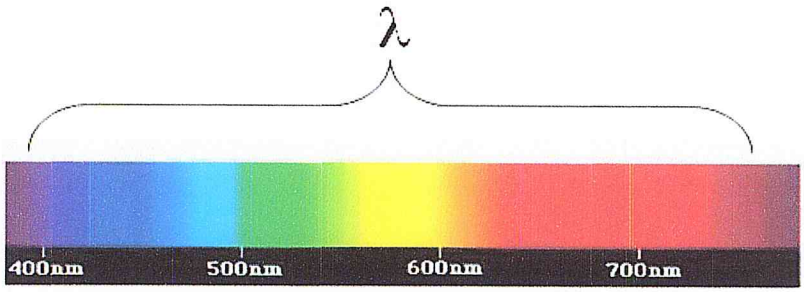


FIGURE 4.11 – longueur d'onde.

- **La normalisation des images LMS :** Pour normaliser l'image LMS, il suffit d'appliquer les équations suivantes à $image_{LMS}$:

$image_{LMS} = image_{LMS} - minimum(image_{LMS})$.

Puis le logarithme est utilisé pour obtenir une normalisation non linéaire $rnonlinear$:

$rnonlinear(z) = \frac{(\log(z+c)-\log(c))}{(\log(1+c)-\log(c))}$

Où c est une valeur alétoir positive avec $c < 1$ et $r(z)$ c'est la localisation du pixel d'une image LMS.

- **Le filtrage gaussien sur les images LMS :** Pour augmenter la visibilité des objets ou autres détails présents sur l'image traitée nous effectuons un filtrage Gaussien sur l'image LMS qui permet d'éliminer les détails de haute fréquence qui contiennent toujours du bruit. L'équation du filtre gaussien est donnée comme suit[38] :

$$G(x,y) = \frac{1}{2\pi\sigma^2} \exp^{-} \frac{(x^2+y^2)}{2\sigma^2} \tag{4.23}$$

Soit $G(x,y)$ le gradient de l'image LMS au point (x,y), le paramètre σ représente la déviation standard. Il faut noter que plus le σ est grand, plus le noyau Gaussien est large et plus le flou appliqué à l'image sera marqué comme apparait dans Figure 4.12 :

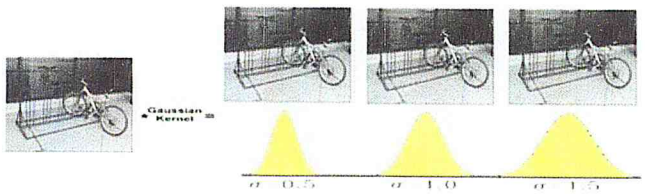


FIGURE 4.12 – Filtrage Gaussien avec différente valeur de σ

Les propriétés de réduction de bruit des filtres Gaussiens peuvent être utilisées en combinaisons avec d'autres filtres comme les filtres Laplaciens (cf. Figure 4.13). On peut par exemple choisir d'appliquer d'abord un filtre Gaussien pour réduire le bruit, avant d'appliquer un filtre Laplacien pour détecter les points autour des quels les variations de luminosité sont importantes. Le filtre Laplacien [2] est définit comme suit :

$$\nabla^2 G(r) = \frac{-1}{\pi\sigma^4}(1 - \frac{r^2}{2\sigma^2}) \exp(-\frac{r^2}{2\sigma^2}) \tag{4.24}$$

Avec $\nabla(r)$ le Laplacien de l'image LMS au point (x,y) et r les point (x,y) .

FIGURE 4.13 – les Filtres gaussien et la placien

Donc après les filtrages Gaussien et Laplacien, nous obtenons une matrice des caractéristiques des composants important de l'image LMS.

2. http ://www.math.unicaen.fr/ reyssat/laplace/laplacien/laplacien.pdf

- **L'analyse en composantes indépendantes (ICA) :** Dans cette étape nous avons utilisé une analyse en composantes indépendantes (ICA) en anglais (Independent Component Analysis) pour extraire les caractéristiques des images LMS dans une scène avec les statistiques naturelles [123],[109], c'est une technique de statistique qui représente une valeur aléatoire multidimensionnelle comme combinaison linéaire des variable arbitraire (composant indépendant) , nous pouvons dire que l'ICA est l'art d'intégrer l'information non Gaussienne dans la recherche des composantes indépendants telle qu'illustré dans Figure 4.14 :

FIGURE 4.14 – Image non Gaussienne

L'ICA est appliquée sur des patches aléatoires de l'image non Gaussienne. Les composantes indépendantes dans un même patch sont calculées par la transformation linéaire suivante :

$$I(x,y) = \sum_i a_i c_i(x,y) \tag{4.25}$$

Telle que (x,y) patch de l'image (image non Gaussien) ai amplitude de la composante indépendante $c_i(x,y)$, elle est calculée comme suit :

$$a_i) = \sum f_i(x,y) I(x,y) \tag{4.26}$$

Si tous les pixels du patch $(x,y)_{nonGaussien}$ sont concaténée dans un simple vecteur I et toutes les amplitudes ai dans un vecteur a, alors l'équation 4.31 devient :

$$I = mci * a \tag{4.27}$$

Avec mci une matrice des composantes indépendantes. Un exemple est présenté dans la Figure 4.15.

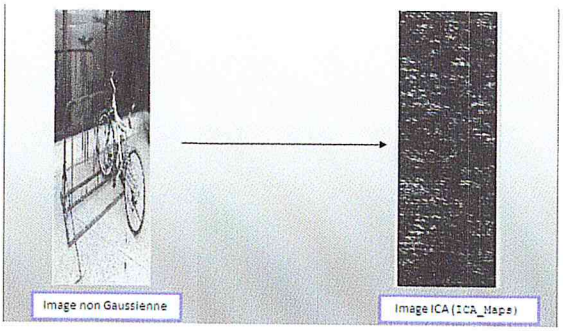


FIGURE 4.15 – Transformation linéaire d'une image non Gaussienne.

La fonction linéaire suivante présente le filtrage des composantes indépendantes.

$$a = Mfic * I \tag{4.28}$$

Avec $Mfic$ une matrice qui contient les composantes indépendantes filtrées. Un exemple est illustré dans la Figure 4.16 :

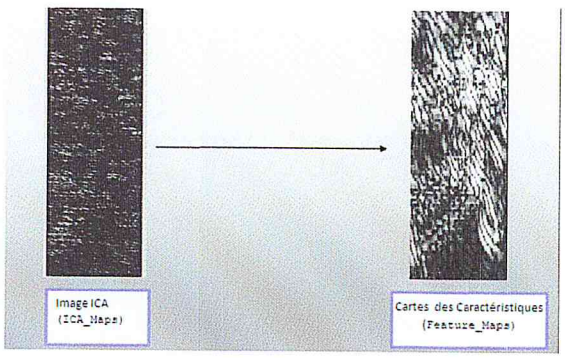


FIGURE 4.16 – Le filtrage des composantes indépendantes.

- **Extraire les cartes de saillance :** La matrice des caractéristiques des composantes indépendantes filtrées est utilisée pour extraire d'autres caractéristiques afin de produire les cartes de saillance par l'utilisation d'un modèle Botton- up [27] (cf. Figure 4.17), c'est un modèle probabiliste qui est calculée comme suit :

$$p(F)^{-1} = F \tag{4.29}$$

F Vecteur des caractéristiques ICA .

Le modèle $p(f)$ est un produit des distributions unidimensionnelles calculé comme suit :

$$p(F = f) = \prod_i p(f_i) \tag{4.30}$$

f_i $i^{éme}$ élément de vecteur F. Telle que (f_i) est calculée par la distribution Gaussienne généralisée (DGG) qui permet de modélisé chacune de ces distributions unidimensionnelles comme suit :

$$p(f_i) = \frac{\theta_1}{2\sigma_i \Gamma(\theta_{i-1})} \exp(-|\frac{f_i}{\gamma_i}|^{\theta}_i \tag{4.31}$$

$\theta_1 > 0$ paramètre de la forme.

$\sigma_i > 0$ paramètre d'échelle.

Γ Fonction gamma.

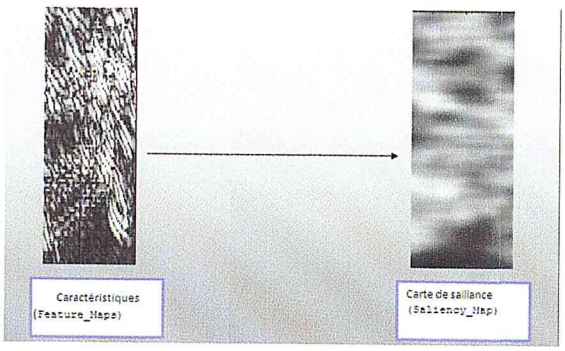


FIGURE 4.17 – les caractéristiques des cartes de saillance

Finalement, nous obtenons un ensemble de patchs avec des caractéristiques de saillance, donc la concaténation de ces patchs donne une image appelée carte de saillance (saliency map) où les objets pertinents sont mis en valeur. Les figures : Figure (a), Figure (b), Figure (c), Figure (d), Figure (e), Figure (f),Figure (g) montrent des exemples de cartes de saillances obtenues sur une sélection d'images.

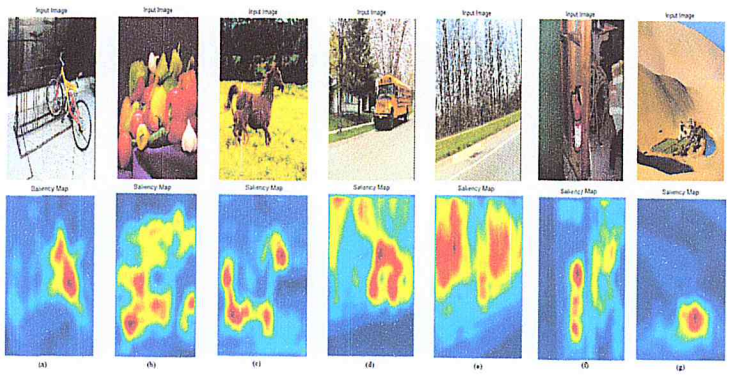


FIGURE 4.18 – de haut en bas :image initiale , cartes de saillances correspondante.

Passage de la carte de saillance à la courbe d'attention pour extraire les images clés

Afin de détecter les images clés de la vidéo, nous avons extrait, pour chaque plan, les cartes de saillances. Parmi les images du même plan, l'image qui contient la valeur d'intensité (attention value) de la carte de saillance la plus élevée est considérée comme image clé (cf. Figure 4.19).

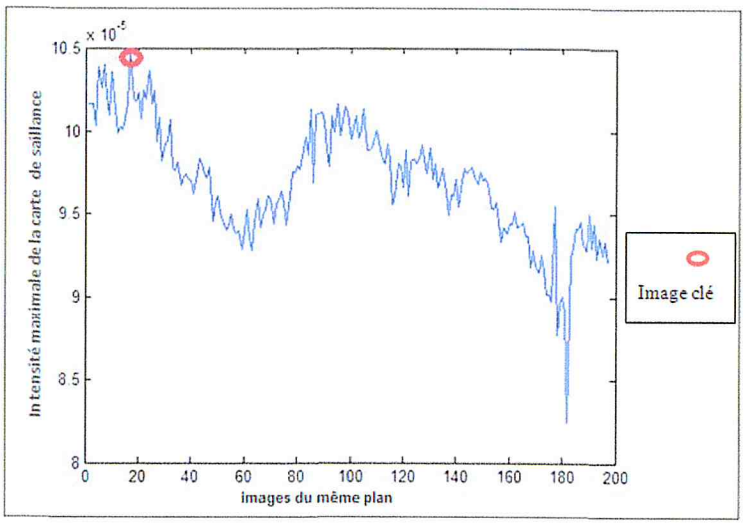


FIGURE 4.19 – Courbe des intensités maximales des cartes de saillances des images appartenant au même plan.

4.2.5 Résultats et interprétations

Afin d'évaluer les performances de l'approche que nous avons proposée pour l'extraction des images clés à partir d'une vidéo, les avons testés sur une sélection de vidéos de la base de données Vidéo SUMMarization (VSUMM). Les résultats obtenus sont présentés dans le Tableau 4.2.

vidéo	seuil	plans	images clé	Précision	Recall	F-mesure
v22	0.49	197 287 365 1769 2118	17 204 358 1748 1820	0,80	0,80	0,8
v47	0.65	277 463 696 998 2160	84 463 488 733 1087	0,71	0,71	0,71
v48	0.3	644 1237 1644 1979 2765 3166 3705	158 686 1320 1875 2765 2803 3184	0,80	0,80	0,80
v50	0.5	703 1739 2144 2746 3244 3824	329 1504 1972 2474 3189 3748 3938	0,62	0,71	0,779
v56	0.75	466 1139 1254 1352 1462 2325	41 469 1196 1352 1391 1655	0,83	0,63	0,716
v57	0.7	1574 2252 3450	1049 1870 2771	1	0,60	0,75
v62	0.87	188 248 630 738 2616	31 213 265 664 822	0,80	1,00	0,88
v66	0.6	128 416 770 983 1063 1305 1696 1843 2187	44 345 575 822 1042 1304 1467 1840 1848	0,78	0,88	0,826
v69	0.36	605 1156 278 2665 2960 3618	13 827 1489 2614 2908 3611	1	0,67	0,8
v70	0.69	364 554 1124 1127 1407	6 366 966 1125 1192	0,75	0,60	0,66

TABLE 4.2 – les mesures d'évaluation de l'approche proposée.

La figure 4.20 représente les valeurs de F-mesure obtenues.

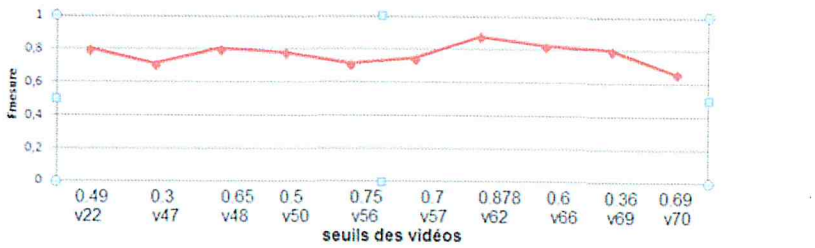


FIGURE 4.20 – Les valeurs de F-mesures pour l'approche proposée.

Les résultats obtenus dans le Tableau 4.2 et la Figure 4.21 montrent que la valeur maximale de F-mesure de l'approche proposée pour l'extraction des image clés est égale à 0.88 et sa valeur moyenne est de 0.77 et le reste des valeurs sont toutes superieures à 0 .66. Nous remarquons qu'il y a une nette amélioration par rapport à la méthode d'extraction des images clés proposée par Poonam S.Jadhav[47]. L'approche proposée est plus performante que celle de de Poonam S.Jadhav[47], la Figure 4.21 montre que les F-mesures de notre approche sont plus élevés que ceux de Poonam S.Jadhav[47].

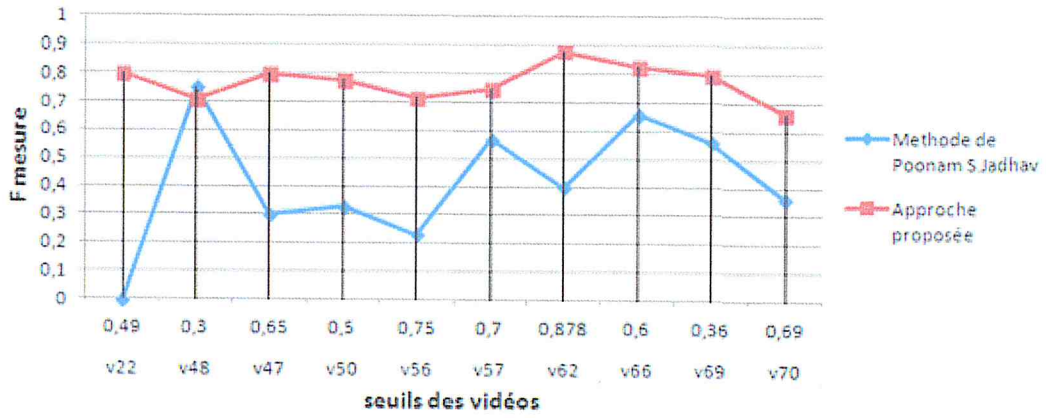


FIGURE 4.21 – Comparaison entre la méthode de Poonam S.Jadhav [47] et l'approche proposée .

4.3 La présentation de l'application

Après avoir présenté, les différents concepts théoriques et pratiques des résumés vidéo, nous nous intéressons dans cette section à la présentation de l'application réalisée où une description de l'environnement matériel et logiciel utilisé seront données.

4.3.1 Environnement matériel

Notre application va être réalisée sur une machine qui comporte les caractéristiques suivantes :

- Marque : HP.
- Modèle : HP 630 notebook PC.
- Processeur : Intel(R) Core (TM) i3-2310M CPU @2 ,10GHZ (2CPUs) .
- RAM : 4GO.
- Système d'exploitation : Windows 7 64bit.

4.3.2 Environnement Logiciel

MATLAB est une abréviation de Matrix LABoratory. Écrit à l'origine en Fortran, par C. Moler, MATLAB était destiné à faciliter l'accès au logiciel matriciel .Sa disponibilité est assurée sur plusieurs plateformes : Sun, Bull, HP, IBM, compatibles PC (DOS, Unix ou Windows), Macintosh, iMac et plusieurs machines parallèles. MATLAB est un environnement puissant, complet et facile à utiliser destiné au calcul scientifique. Il possède les particularités suivantes :

- Puissance de calcul.
- La continuité parmi les valeurs entières, réelle et complexes.
- La compréhension de la bibliothèque mathématique.
- L'inclusion des fonctions d'interface graphique et des unitaires dans l'outil graphique.
- La possibilité de liaison avec les autres langages classiques de programmations(c ou fortran).

Dans MATLAB, aucune déclaration n'est à effectuer sur les nombres .En effet, il n'existe pas de distinction entre les nombres entiers, les nombres réels, les nombre complexes, et la simple ou double précision .cette caractéristique rend le mode de programmation très facile et très rapide. La bibliothèque des fonctions mathématiques d ans MATLAB donne des analyses mathématiques très simples. En effet, l'utilisateur peut exécuter dans le mode commande n'importe quelle fonction mathématique se trouvant dans la bibliothèque sans avoir à recourir à recourir à la programmation Pour l'interface graphique, des représentations scientifiques et même artistiques des objets peuvent être crées sur l'écran en utilisant les expressions mathématiques .Les graphiques sur Matlab sont simples et attirent l'attention des utilisateurs, vu les possibilités importantes offertes par ce logiciel [3]

4.3.3 Les captures d'écran du logiciel développé

La fenêtre principale de notre application est présentée dans la Figure 4.22.

FIGURE 4.22 – L'interface d'accueil.

Le bouton « START » permet de visualiser une autre interface comme le montre la Figure 4.23 :

3. http ://www.iro.umontreal.ca/mignotte/IFT2425/Matlab.pdf

FIGURE 4.23 – Le traitement sur la vidéo.

Le mot en ligne signifie que l'affichage des résultats sera au même temps que l'exécution de programme.

Le bouton « ouvrir » permet de donner la main pour choisir une vidéo et la visualiser comme illustré dans les figures 4.24.

FIGURE 4.24 – La sélection de la vidéo.

Le bouton « AFFICHER» permet de visualiser une autre interface comme le montre la Figure 4.25

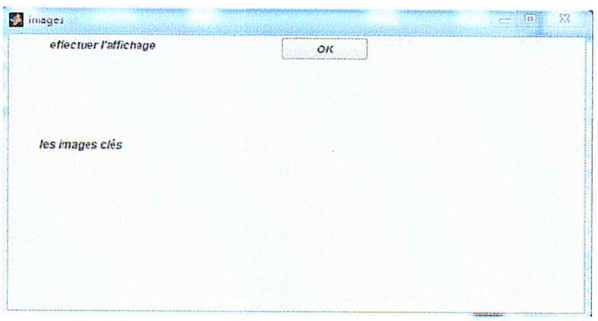


FIGURE 4.25 – L'espace d'affichage des images clés de la vidéo sélectionnée.

Le bouton « ok » permet de commencer l'affichage des transitions brusques et les images

clés de la vidéo sélectionnée. L'affichage des images clés peut prendre beaucoup de temps (en fonction de la machine), comme illustré dans la figure 4.26 :

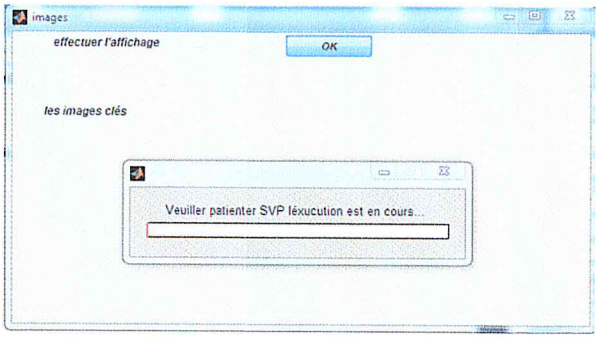


FIGURE 4.26 – wait bar.

Lorsqu'on appuie sur le bouton « ok », une 'Wait bar' indique que le traitement de vidéo est en cours.

La figure 4.27 représente la lecteur de la vidéo intégrée . La figure 4.28 représente les

FIGURE 4.27 – la lecture d'une vidéo.

images clés de la vidéo sélectionnée :

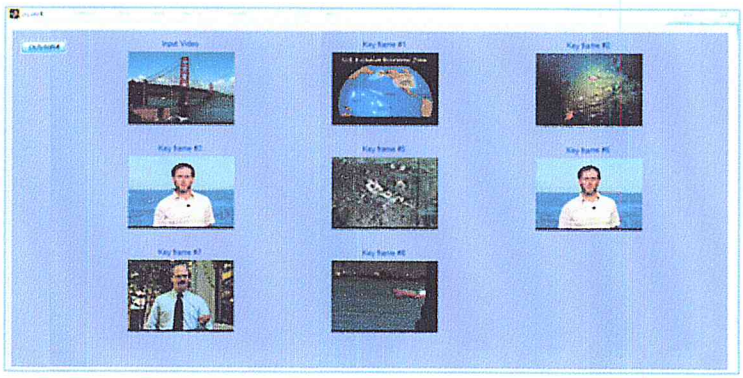


FIGURE 4.28 – les images clés d'une vidéo.

4.4 Conclusion

Dans ce chapitre nous avons présenté une nouvelle approche pour l'extraction des images clés, cette approche utilise les transitions brusques ainsi que les cartes de saillances pour la sélection des images clés à partir d'une vidéo. La robustesse de notre approche a été démontrée par des mesures statistiques ainsi qu'avec une comparaison avec la méthode de S.Jadhav ,C.O Conaire [16][47].

Conclusion et perspectives

Conclusion

Durant ce mémoire, nous nous sommes intéressés à la création de résumé de vidéo. L'objectif de notre travail est d'extraire les images les plus représentatives du contenu de la vidéo afin de donner un aperçu rapide et succinct de son contenu. Le résumé vidéo peut également être utilisé dans de nombreuses applications comme la classification, la recherche par exemple et la navigation dans une base de vidéos.

Nous avons proposé une nouvelle méthode de génération de résumé vidéo qui repose sur des informations d'attention visuelles extraites à partir des images de la vidéo. Notre étude montre que les systèmes à base d'attention visuelle sont plus efficaces que les systèmes à base de caractéristiques de bas niveau. Le schéma proposé extrait efficacement les parties saillantes des images vidéo. Les résultats expérimentaux basés sur plusieurs critères d'évaluation indiquent que le schéma proposé extraits les images clés sémantiquement pertinentes par rapport aux données de la base de donnée VSUMM.

4.5 Perspectives

La poursuite de ce travail de mémoire à court terme va concerner la fusion de plus de caractéristiques pour la génération de résumé vidéo telles que les caractéristiques du mouvement, les caractéristiques couleurs, avec ceux de l'attention visuelle. Cependant, d'autres caractéristiques pourraient être considérées comme la bande son ou le texte. Par exemple, la bande son apporte de nombreuses informations de haut niveau sur le contenu des vidéos. Une étude sur la combinaison des caractéristiques visuelles et sonores pourrait enrichir les méthodes de résumé et toutes les applications qui en découlent.

Bibliographie

[1] Abdessalem Ben Abdelali, Abdeellatif Mtibaa, Elbey Bourennane, and Mohamed Abid. Etude du descripteur structure de couleurs pour la segmentation temporelle de la vidéo vers une implantation temps réel. 2007.

[2] Akihito Akutsu, Yoshinobu Tonomura, Hideo Hashimoto, and Yuji Ohba. Video indexing using motion vectors. In *Applications in Optical Science and Engineering*, pages 1522–1530. International Society for Optics and Photonics, 1992.

[3] Adnan M Alattar. Detecting and compressing dissolve regions in video sequences with a dvi multimedia image compression algorithm. In *Circuits and Systems, 1993., ISCAS'93, 1993 IEEE International Symposium on*, pages 13–16. IEEE, 1993.

[4] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology : tool for the unification of biology. *Nature genetics*, 25(1) :25–29, 2000.

[5] Jacques Aumont. *L'image*. Armand Colin, 2011.

[6] Alan Baddeley. Working memory. *Science*, 255(5044) :556–559, 1992.

[7] Stanley T Birchfield and Sriram Rangarajan. Spatiograms versus histograms for region-based tracking. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 1158–1163. IEEE, 2005.

[8] Stanley T Birchfield and Sriram Rangarajan. Spatiograms versus histograms for region-based tracking. In *Computer Vision and Pattern Recognition, 2005.*

CVPR 2005. IEEE Computer Society Conference on, volume 2, pages 1158–1163. IEEE, 2005.

[9] Christophe Bordeux, Ronan Boulic, and Daniel Thalmann. An efficient and flexible perception pipeline for autonomous agents. In *Computer Graphics Forum*, volume 18, pages 23–30. Wiley Online Library, 1999.

[10] Lionel Brunel. *Indexation vidéo par l'analyse de codage.* PhD thesis, Université Nice Sophia Antipolis, 2004.

[11] Roberto Brunelli, Ornella Mich, and Carla Maria Modena. A survey on the automatic indexing of video data. *Journal of visual communication and image representation*, 10(2) :78–112, 1999.

[12] Cheng Cai, Kin Man Lam, and Zheng Tan. An efficient scene break detection based on linear prediction. In *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*, pages 555–558. IEEE, 2004.

[13] Janko Calic and BT Thomas. Spatial analysis in key-frame extraction using video segmentation. In *Workshop on Image Analysis for Multimedia Interactive Services*, 2004.

[14] Sylvain Chevallier and Philippe Tarroux. Implémentation d'un mécanisme de" covert attention" avec un réseau de neurones impulsionnels. In *Deuxième conférence française de Neurosciences Computationnelles," Neurocomp08"*, 2008.

[15] Dorin Comaniciu and Visvanathan Ramesh. Real-time tracking of non-rigid objects using mean shift, July 8 2003. US Patent 6,590,999.

[16] Ciarán O Conaire, Noel E O'Connor, and Alan F Smeaton. An improved spatiogram similarity measure for robust object localisation. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–1069. IEEE, 2007.

[17] Edmond Couchot and Norbert Hillaire. *L'art numérique.* Flammarion, 2003.

[18] Nicolas Courty, Eric Marchand, and Bruno Arnaldi. A new application for saliency maps : Synthetic vision of autonomous actors. In *Image Processing,*

2003. ICIP 2003. Proceedings. 2003 International Conference on, volume 3, pages III–1065. IEEE, 2003.

[19] Gilles Deleuze. *Cinéma 1-L'image-mouvement.* Minuit, 2014.

[20] Cédric Demonceaux. *Etude du mouvement dans les séquences d'images par analyse d'ondelettes et modélisation markovienne hiérarchique. Application à la détection d'obstacles dans un milieu routier.* PhD thesis, Université de Picardie Jules Verne, 2004.

[21] Antoine Denis. Travaux pratiques de télédétection spatiale. 2012.

[22] Mr Sandip T Dhagdi and Dr PR Deshmukh. Keyframe based video summarization using automatic threshold & edge matching rate. *International Journal of Scientific and Research Publications*, 2(7) :1–12, 2012.

[23] Gerald M Edelman. *La science du cerveau et la connaissance.* Odile Jacob, 2007.

[24] WAC Fernando, Cedric N Canagarajah, and David R Bull. Video segmentation and classification for content-based storage and retrieval using motion vectors. In *Electronic Imaging'99*, pages 687–698. International Society for Optics and Photonics, 1998.

[25] Warnakulasuriya Anil Chandana Fernando, Cedric Nishan Canagarajah, and David R Bull. A unified approach to scene change detection in uncompressed and compressed video. *Consumer Electronics, IEEE Transactions on*, 46(3) :769–779, 2000.

[26] Steven J Friedman, Karen A Hargrove, Joseph M Joy, Nathan P Myhrvold, Sunita Shrivastava, and Gideon A Yuval. For approximating a high color resolution image with a low one, October 3 1995. US Patent 5,455,600.

[27] Dashan Gao and Nuno Vasconcelos. Bottom-up saliency is a discriminant process. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–6. IEEE, 2007.

[28] Wen Gao, Yonghong Tian, Tiejun Huang, and Qiang Yang. Vlogging : A survey of videoblogging technology on the web. *ACM Computing Surveys (CSUR)*, 42(4) :15, 2010.

[29] Ullas Gargi, Rangachar Kasturi, and Susan H Strayer. Performance characterization of video-shot-change detection methods. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(1) :1–13, 2000.

[30] Syntyche Gbèhounou, François Lecellier, and Christine Fernandez-Maloigne. Extraction de l'impact émotionnel des images. *Traitement de Signal*, pages 409–432, 2012.

[31] Pierre Gouzerh and Michel Che. From scheele and berzelius to müller : Polyoxometalates (poms) revisited and the missing link between the bottom up and top down approaches. *L'actualite chimique*, (298) :9–22, 2006.

[32] Mickael Guironnet. *Méthodes de résumé de vidéo à partir d'informations bas niveau, du mouvement de caméra ou de l'attention visuelle.* PhD thesis, Université Joseph-Fourier-Grenoble I, 2006.

[33] Gaëlle Hallair. Vidéo et pratique de la géographie. *EchoGéo*, (2), 2007.

[34] Allan Hanbury. *Morphologie Mathématique sur le Cercle Unité, avec applications aux teintes et aux textures orientées.* PhD thesis, École Nationale Supérieure des Mines de Paris, 2002.

[35] Alan Hanjalic. Shot-boundary detection : unraveled and resolved ? *Circuits and Systems for Video Technology, IEEE Transactions on*, 12(2) :90–105, 2002.

[36] Mary M Hayhoe, Dana H Ballard, Jochen Triesch, Hiroyuki Shinoda, Pilar Aivar, and Brian Sullivan. Vision in natural and virtual environments. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 7–13. ACM, 2002.

[37] John M Henderson and Andrew Hollingworth. Eye movements and visual memory : Detecting changes to saccade targets in scenes. *Perception & Psychophysics*, 65(1) :58–71, 2003.

[38] Walid Hizem. *Capteur intelligent pour la reconnaissance de visage.* PhD thesis, Evry, Institut national des télécommunications, 2009.

[39] Jérémie Hornus and Patrick Andries. Rendu de polices sur écrans à cristaux liquides. *Document numérique*, 9(3) :87–116, 2007.

[40] Fayez Idris and Sethuraman Panchanathan. Review of image and video indexing techniques. *Journal of visual communication and image representation*, 8(2) :146–166, 1997.

[41] Fayez Idris and Sethuraman Panchanathan. Review of image and video indexing techniques. *Journal of visual communication and image representation*, 8(2) :146–166, 1997.

[42] Laurent Itti, Nitin Dhavale, and Frederic Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Optical science and technology, SPIE's 48th annual meeting*, pages 64–78. International Society for Optics and Photonics, 2004.

[43] Laurent Itti and Christof Koch. Learning to detect salient objects in natural scenes using visual attention. In *Image Understanding Workshop*, pages 1201–1206. Citeseer, 1999.

[44] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10) :1489–1506, 2000.

[45] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3) :194–203, 2001.

[46] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11) :1254–1259, 1998.

[47] Mrs Poonam S Jadhav and Dipti S Jadhav. Video summarization using higher order color moments (vsuhcm). *Procedia Computer Science*, 45 :275–281, 2015.

[48] Shanon X Ju, Michael J Black, Scott Minneman, and Don Kimber. Summarization of videotaped presentations : automatic analysis of motion and gesture. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5) :686–696, 1998.

[49] Kathleen Julié and Laurent Perrot. *Enseigner l'anglais*. Hachette éducation, 2008.

[50] Christopher Kanan and Garrison Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *Computer Vision and*

Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 2472–2479. IEEE, 2010.

[51] Cécile Kattnig. *Gestion et diffusion d'un fond d'image.* Armand Colin, 2005.

[52] Nazih Khaddaj Mallat, Emilia Moldovan, and Serioja O Tatu. Comparative demodulation results for six-port and conventional 60 ghz direct conversion receivers. *Progress In Electromagnetics Research*, 84 :437–449, 2008.

[53] Ewa Kijak. *Structuration multimodale des vidéos de sport par modèles stochastiques.* PhD thesis, Université Rennes 1, 2003.

[54] Takeshi Kikukawa and Satomi Kawafuchi. Development of an automatic summary editing system for the audio-visual resources. *Trans. Inst. Electron., Inform., Commun. Eng*, 75 :204–212, 1992.

[55] Helmut Kraus and Uwe Steinmueller. *Mastering HD video with your DSLR.* Rocky Nook, Inc., 2010.

[56] James J Kuffner Jr and Jean-Claude Latombe. Fast synthetic vision, memory, and learning models for virtual humans. In *Computer Animation, 1999. Proceedings*, pages 118–127. IEEE, 1999.

[57] Reginald L Lagendijk, Alan Hanjalic, Marco Ceccarelli, Mario Soletic, and Eric Persoon. Visual search in a smash system. In *Image Processing, 1996. Proceedings., International Conference on*, volume 3, pages 671–674. IEEE, 1996.

[58] Saadi Lahlou. L'activité du point de vue de l'acteur et la question de l'intersubjectivité [huit années d'expériences avec des caméras miniutarisées fixées au front des acteurs (subcams)]. *Communications*, 80(1) :209–234, 2006.

[59] Rachida LAKHDARI. La détection des micros calcifications dans l'image mammographie. 2011.

[60] Frédéric Landragin. Modélisation de la saillance visuelle et linguistique. In *Sixième Colloque des Jeunes Chercheurs en Sciences Cognitives (CJCSC'05)*, pages 157–162. Université Bordeaux 2, 2005.

[61] Olivier Le Meur. *Attention sélective en visualisation d'images fixes et animées affichées sur écran : modèles et évaluation de performances-application.* PhD thesis, Nantes, 2005.

[62] Sébastien Lefevre. *Détection d'événements dans une séquence vidéo.* PhD thesis, Université François Rabelais-Tours, 2002.

[63] LT Les matrices MegaPower. dôme sensornet ou ad bi-directionnel" via le câble coaxial"(utc/up-the-coax).• nouvelle fonction.

[64] Hervé Liebgott. *Synthèse de réponse impulsionnelle en imagerie ultrasonore pour l'estimation vectorielle du deplacement.* PhD thesis, Villeurbanne, INSA, 2005.

[65] Rainer Lienhart. Reliable transition detection in videos : A survey and practitioner's guide. *International journal of image and graphics*, 1(03) :469–486, 2001.

[66] Rainer W Lienhart. Comparison of automatic shot boundary detection algorithms. In *Electronic Imaging'99*, pages 290–301. International Society for Optics and Photonics, 1998.

[67] Dokshin Lim. *Modélisation du processus de conception centrée utilisateur, basée sur l'intégration des méthodes et outils de l'ergonomie cognitive : application à la conception d'IHM pour la télévision interactive.* PhD thesis, Arts et Métiers ParisTech, 2003.

[68] Tianming Liu, Hong-Jiang Zhang, and Feihu Qi. A novel video key-frame-extraction algorithm based on perceived motion energy model. *Circuits and Systems for Video Technology, IEEE Transactions on*, 13(10) :1006–1013, 2003.

[69] HB Lu, YJ Zhang, and YR Yao. Robust gradual scene change detection. In *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, volume 3, pages 304–308. IEEE, 1999.

[70] G Lupatini, Caterina Saraceno, and Riccardo Leonardi. Scene break detection : a comparison. In *Research Issues In Data Engineering, 1998.'Continuous-Media Databases and Applications'. Proceedings., Eighth International Workshop on*, pages 34–41. IEEE, 1998.

[71] A Manoury and H Nicolas. Segmentation temporelle de vidéos numériques fondée sur l'utilisation de mosaïques 1d. In *19 Colloque sur le traitement du*

signal et des images, FRA, 2003. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 2003.

[72] Markos Mentzelopoulos and Alexandra Psarrou. Key-frame extraction algorithm using entropy difference. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 39–45. ACM, 2004.

[73] Maurice Merleau-Ponty. *Phénoménologie de la perception.* éditions Gallimard, 2013.

[74] Kathleen Mullaniff. Fade in fade out. 2012.

[75] Akio Nagasaka and Yuzuru Tanaka. Automatic video indexing and full-video search for object appearances. 1992.

[76] Ken Nakayama and Manfred Mackeben. Sustained and transient components of focal visual attention. *Vision research*, 29(11) :1631–1647, 1989.

[77] Jeho Nam and Ahmed H Tewfik. Dissolve transition detection using b-splines interpolation. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 3, pages 1349–1352. IEEE, 2000.

[78] Vidhya Navalpakkam and Laurent Itti. A goal oriented attention guidance model. In *Biologically motivated computer vision*, pages 453–461. Springer, 2002.

[79] Chong-Wah Ngo, Ting-Chuen Pong, and Roland T Chin. Detection of gradual transitions through temporal slice analysis. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 1. IEEE, 1999.

[80] Chong-Wah Ngo, Ting-Chuen Pong, and Roland T Chin. Video partitioning by temporal slice coherency. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(8) :941–953, 2001.

[81] Hansrudi Noser, Olivier Renault, Daniel Thalmann, and Nadia Magnenat Thalmann. Navigation for digital actors based on synthetic vision, memory, and learning. *Computers & graphics*, 19(1) :7–19, 1995.

[82] Atrde Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. Top-down control of visual attention in object detection. In *Image processing,*

2003. icip 2003. proceedings. 2003 international conference on, volume 1, pages I–253. IEEE, 2003.

[83] Kintu Patel Oriental. Key frame extraction based on block based histogram difference and edge matching rate.

[84] C Alejandro Parraga, J Vazquez-Corral, and Maria Vanrell. A new cone activation-based natural images dataset. *Perception*, 36 :180, 2009.

[85] Stéphane Péchard. *Qualité d'usage en télévision haute définition : évaluations subjectives et métriques objectives.* PhD thesis, Université de Nantes, 2008.

[86] Christopher Peters and Carol O'Sullivan. Bottom-up visual attention for virtual human animation. In *Computer Animation and Social Agents, 2003. 16th International Conference on*, pages 111–117. IEEE, 2003.

[87] S Pettigrand et al. Mesures 3d de topographies et de vibrations a l'echelle (sub) micrometrique par microscopie optique interferometrique. In *Proc. Club CMOI, Methodes et Techniques Optiques pour l'Industrie*, 2002.

[88] Shumeet Baluja Dean A Pomerleau. Using a saliency map for active spatial selective attention : Implementation & initial results. *Advances in Neural Information Processing Systems 7*, 7 :451, 1995.

[89] Sarah V Porter, Majid Mirmehdi, and Barry T Thomas. Detection and classification of shot transitions. In *BMVC*, pages 1–10, 2001.

[90] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. Numerical recipes (cambridge, 1992.

[91] Christophe Reffay, Thierry Chanier, Muriel Noras, and Marie-Laure Betbeder. Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. *Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation*, 15 :xx, 2008.

[92] Olivier Renault, Nadia Magnenat Thalmann, and Daniel Thalmann. A vision-based approach to behavioural animation. *The journal of Visualization and computer animation*, 1(1) :18–21, 1990.

[93] Ronald A Rensink, J Kevin O'Regan, and James J Clark. To see or not to see : The need for attention to perceive changes in scenes. *Psychological science*, 8(5) :368–373, 1997.

[94] Craig W Reynolds. Flocks, herds and schools : A distributed behavioral model. In *ACM SIGGRAPH computer graphics*, volume 21, pages 25–34. ACM, 1987.

[95] Derek A Roff and Paul Bentzen. The statistical analysis of mitochondrial dna polymorphisms : chi 2 and the problem of small samples. *Molecular biology and evolution*, 6(5) :539–545, 1989.

[96] Jaiwei Rong, Yu-Fei Ma, and Lide Wu. Gradual transition detection using em curve fitting. In *Multimedia Modelling Conference, 2005. MMM 2005. Proceedings of the 11th International*, pages 364–369. IEEE, 2005.

[97] Edmundo Sáez, José I Benavides, and Nicolas Guil. Combining luminance and edge based metrics for robust temporal video segmentation. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 4, pages 2231–2234. IEEE, 2004.

[98] Emmanuel Schmitt. *Contribution au Système d'Information d'un Produit «Bois». Appariement automatique de pièces de bois selon des critères de couleur et de texture.* PhD thesis, Université Henri Poincaré-Nancy I, 2007.

[99] Behzad Shahraray. Scene change detection and content-based sampling of video sequences. In *IS&T/SPIE's Symposium on Electronic Imaging : Science & Technology*, pages 2–13. International Society for Optics and Photonics, 1995.

[100] Alain Simac. *Modélisation et gestion de concepts, en particulier temporels, pour l'assistance à la caractérisation de séquences d'images.* PhD thesis, Université de Grenoble, 2011.

[101] Daniel J Simons and Ronald A Rensink. Change blindness : Past, present, and future. *Trends in cognitive sciences*, 9(1) :16–20, 2005.

[102] Fabrice Souvannavong. *Indexation et recherche de plans vidéo par le contenu sémantique.* PhD thesis, Télécom ParisTech, 2005.

[103] Yaoru Sun and Robert Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146(1) :77–123, 2003.

[104] Eugene Taylor. *William James on consciousness beyond the margin.* Princeton University Press, 1996.

[105] Demetri Terzopoulos and Tamer E Rabie. Animat vision : Active vision in artificial animals. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 801–808. IEEE, 1995.

[106] Ba Tu Truong, Chitra Dorai, and Svetha Venkatesh. New enhancements to cut, fade, and dissolve detection processes in video segmentation. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 219–227. ACM, 2000.

[107] Shingo Uchihashi, Jonathan Foote, Andreas Girgensohn, and John Boreczky. Video manga : generating semantically meaningful video summaries. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 383–392. ACM, 1999.

[108] Hirotada Ueda, Takafumi Miyatake, and Satoshi Yoshizawa. Impact : an interactive natural-motion-picture dedicated multimedia authoring system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 343–350. ACM, 1991.

[109] J Hans van Hateren and Dan L Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London B : Biological Sciences*, 265(1412) :2315–2320, 1998.

[110] Nicolas Vandenbroucke. *Segmentation d'images couleur par classification de pixels dans des espaces d'attributs colorimétriques adaptés. Application à l'analyse d'images de football.* PhD thesis, université de Rouen, 2000.

[111] Morgan Veyret. Focalisation de l'attention visuelle. 2004.

[112] Morgan Veyret and Eric Maisel. Attention-based target tracking for an augmented reality application. 2006.

[113] Theodore Vlachos. Cut detection in video sequences using phase correlation. *Signal Processing Letters, IEEE*, 7(7) :173–175, 2000.

[114] Anthony Whitehead, Prosenjit Bose, and Robert Laganiere. Feature based cut detection with automatic threshold selection. In *Image and Video Retrieval*, pages 410–418. Springer, 2004.

[115] Wayne Wolf. Key frame selection by motion analysis. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pages 1228–1231. IEEE, 1996.

[116] Itheri Yahiaoui. *Construction automatique de résumés vidéos.* PhD thesis, Paris, ENST, 2003.

[117] Jek Charlson So Yu, Mohan S Kankanhalli, and P Mulhen. Semantic video summarization in compressed domain mpeg video. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 3, pages III–329. IEEE, 2003.

[118] Yusseri Yusoff, W Christmas, and Josef Kittler. A study on automatic shot change detection. In *Multimedia Applications, Services and Techniques—ECMAST'98*, pages 177–189. Springer, 1998.

[119] Ramin Zabih, Justin Miller, and Kevin Mai. A feature-based algorithm for detecting and classifying production effects. *Multimedia systems*, 7(2) :119–128, 1999.

[120] Ramin Zabih, Justin Miller, and Kevin Mai. A feature-based algorithm for detecting and classifying production effects. *Multimedia systems*, 7(2) :119–128, 1999.

[121] Hong Jiang Zhang, Jianhua Wu, Di Zhong, and Stephen W Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern recognition*, 30(4) :643–658, 1997.

[122] HongJiang Zhang, Atreyi Kankanhalli, and Stephen W Smoliar. Automatic partitioning of full-motion video. *Multimedia systems*, 1(1) :10–28, 1993.

[123] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun : A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7) :32–32, 2008.

[124] Jian Zhou and Xiao-Ping Zhang. Video shot boundary detection using independent component analysis. In *ICASSP (2)*, pages 541–544, 2005.

[125] Yueting Zhuang, Yong Rui, Thomas S Huang, and Sharad Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *Image Processing, 1998.*

ICIP 98. Proceedings. 1998 International Conference on, volume 1, pages 866–870. IEEE, 1998.

[126] Zoran Zivkovic and Ben Krose. An em-like algorithm for color-histogram-based object tracking. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–798. IEEE, 2004.