Stage M2

État de l'art sur les méthodes de détermination d'une image clé d'une séquence vidéo

Benjamin Serva
Master 2 IMAGINE
Université de Montpellier

21 mars 2025

Encadrants :

Olivier Strauss & William Puech & Frédéric Comby

Contents

1 Mise en contexte

L'objectif de cette état de l'art est de mettre évidence les différentes méthodes existantes qui permettent de déterminer une image clé/représentative d'une vidéo. Dans notre cas c'est un extrait considéré comme un plan.

2 Méthodes par segmentation (détection de plan)

Une grande partie des papiers se base sur cette méthode notamment dans [1], où la méthode consiste à faire une segmentation par plan de la vidéo. Ainsi pour chaque plan détecté on garde la première et la dernière frame, ce qui nous donne une liste de N frames clés.

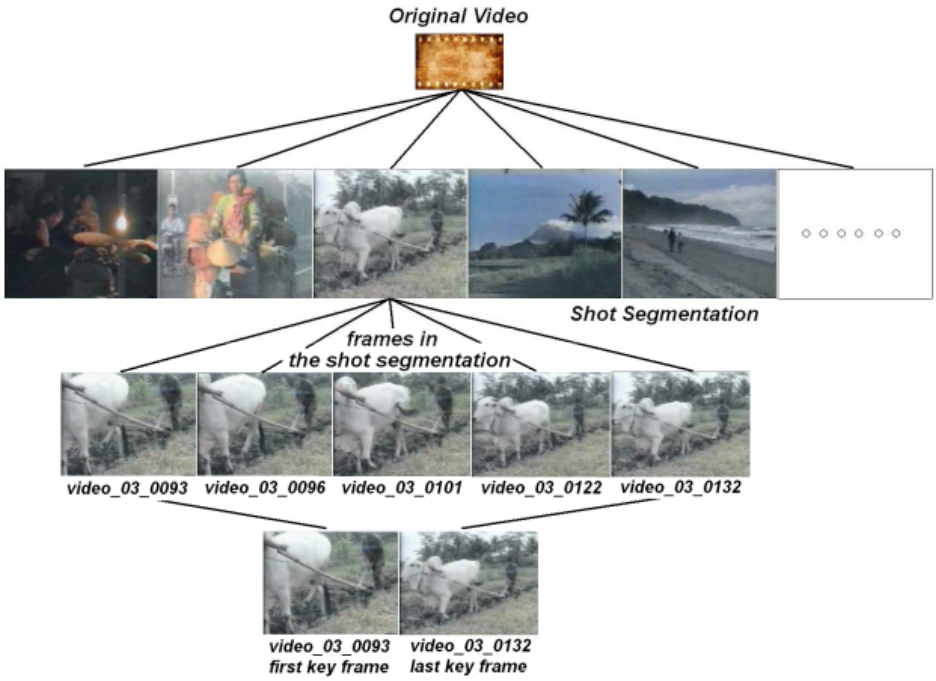

Figure 1: The Structure Video of Key Frame Selection

3 Méthodes de sélection par thèmes

Dans ce papier [2] ils font la distinction entre deux types de vidéos : édités (film, pub, clip etc...) et non édités (vidéo sans post production donc filmé d'un seul trait).

3.1 Première étape : extraction d'images clés

3.1.1 Vidéos non éditées

Différentes méthodes servent à définir des images clés:

- Les différences d'histogramme de couleur et de disposition de couleur sont calculées et comparées à des seuils empiriques.
- Une mise au point après une période de mouvement peut indiquer une scène ou un objet d'intérêt.
- Les objets en mouvement sont détectés.

- Détection de visages.
- Le contenu audio est analysé pour reconnaître des événements tels que le rire ou la parole.

Les images candidates sont regroupées en clusters à l'aide de l'algorithme K-means ou d'une méthode de construction d'ensemble adaptative.
Un score d'importance est calculé pour chaque image clé candidate en fonction des mouvements de caméra, des visages humains, de la taille et de la position des objets en mouvement, ainsi que des événements audio.
La qualité d'image est également évaluée (netteté, luminosité et contraste). Une image représentative est sélectionnée à partir de chaque cluster en fonction du score d'importance, de la proximité avec le centre du cluster et de la qualité d'image.

3.1.2 Vidéo éditées

Tout d'abord la vidéo est segmentée en plan, puis ensuite trois modes de sélections sont possible :

- Une image par plan : sélection de l'image de meilleur qualité.
- Nombre fixe d'images clés prédéterminé.
- Taux fixe d'images clés, sélectionné à intervalle régulier.

3.2 Deuxième étape : classement basé sur le thème

Des mots clés sont obtenus à partir du titre et de la description de la vidéo, ensuite une recherche d'images est effectuée (par exemple via Google images) avec ces mots clés. Grâce à ces résultats un modèle visuel du thème est construit à partir des caractéristiques communes des images obtenues, ce qui va permettre d'affiner la sélection des images clés notamment en se basant sur les couleurs.

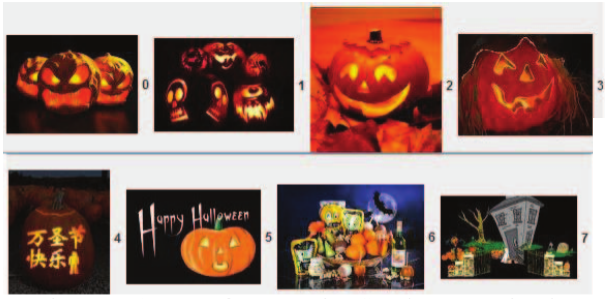

Fig.3. Top 8 returns from querying Google Image using the keyword "Halloween".

Fig.4. Principle color components of the "Halloween" theme.

Figure 2: Exemple de construction de modèle visuel

Fig.5. Keyframes ordered by their distance to the theme model.
Top left is the best, bottom right is the worst.

Figure 3: Image clés sélectionné en fonction du modèle précédent

Fig.6. Keyframes ordered by image quality only.

Fig.7. Keyframes ordered by a study participant.

Figure 4: Images clés par méthodes classique (à gauche) et "valeurs de vérités" (à droite)

4 Méthodes d'extraction d'images clés par représentation parcimonieuse

Ce document [3] propose une approche pour extraire automatiquement des images clés de vidéos en s'appuyant sur la théorie de la représentation parcimonieuse. L'approche se distingue par son indépendance vis-à-vis de la détection de plans ou de la segmentation préalable.

4.1 Première étape : Projection et extraction de caractéristiques

- Chaque image de la vidéo est convertie en un vecteur (avec les pixels réorganisés de manière lexicographique).
- Ce vecteur est projeté dans un espace de dimension réduite grâce à une matrice de projection aléatoire
- Cette projection permet de réduire le coût de calcul par rapport à des descripteurs classiques (SIFT, GIST, histogramme de couleurs, etc.).

4.2 Deuxième étape : Représentation parcimonieuse et construction du dictionnaire

- Chaque vecteur est reconstruit comme une combinaison linéaire parcimonieuse des colonnes d'un dictionnaire surcomplet

- Le vecteur de coefficients est contraint à être non négatif et contient peu d'éléments non nuls.
- L'optimisation se fait via la minimisation de l'erreur de reconstruction combinée à une régularisation.

4.3 Troisième étape : Clustering et sélection des images clés

- À partir des coefficients on construit une matrice symétrique qui capture les similarités entre images.
- Une contrainte temporelle est appliquée pour favoriser l'influence des images proches dans le temps.
- Un clustering normalisé est ensuite appliqué, Le nombre de clusters correspond au nombre désiré d'images clés.
- Pour chaque cluster, les images sont ordonnées temporellement et l'image médiane est retenue comme image clé.

4.4 Résultats et performances

- L'approche a été testée sur une base de 100 vidéos variés.
- Les résultats montrent une bonne correspondance avec la "vérité terrain" établie par des juges humains, et une performance supérieure à une méthode basée sur la détection de mouvement de caméra.
- L'utilisation d'une projection aléatoire et de la représentation parcimonieuse permet une réduction significative du temps de calcul, rendant l'approche particulièrement efficace pour le traitement de grandes quantités de vidéos.

Figure 5: "Valeurs de vérités" de la vidéo

Figure 6: Image clés sélectionné par la méthode

4.5 Conclusions et perspectives

- L'approche proposée permet d'extraire efficacement des images clés sans recourir à des techniques complexes de segmentation ou de détection de plans.
- Les résultats expérimentaux démontrent sa faisabilité et sa robustesse sur des vidéos non structurées.

5 Méthode d'extraction d'images clés par utilisation de l'image epitome

Ce document [4] présente une approche pour extraire automatiquement des images clés à partir de vidéos, en s'appuyant sur la représentation par image epitome. Comme pour la méthode précédente elle ne nécessite aucune segmentation.

5.1 Première étape : Représentation par image epitome

- Pour chaque image de la vidéo, une représentation condensée, appelée "image epitome", est construite. L'epitome regroupe les informations visuelles essentielles (texture, contours, couleurs, etc.) en agrégeant de petits patchs caractéristiques.
- Cette représentation est obtenue en associant de manière probabiliste chaque patch de l'image à une région de l'epitome via un modèle gaussien. Les paramètres (moyennes et variances) des distributions associées à chaque position de l'epitome sont estimés.
- Un algorithme d'espérance-maximisation (EM) est utilisé pour optimiser la correspondance entre les patchs de l'image originale et ceux de l'epitome, garantissant ainsi que cette dernière conserve les aspects les plus saillants de l'image.

5.2 Deuxième étape : Calcul de la dissimilarité entre images

- Chaque image est désormais représentée par son vecteur epitome, qui résume l'information visuelle de manière compacte.
- Pour mesurer la dissimilarité entre deux images, on modélise ces vecteurs par des mélanges de Gaussiennes. La comparaison s'effectue à l'aide d'une divergence de Kullback-Leibler.

5.3 Troisième étape : Sélection des images clés via l'algorithme min-max

- Une fois les dissimilarités calculées pour chaque paire d'images, la méthode procède à la sélection des images clés en appliquant un algorithme min-max.
- La première sélection consiste à identifier les deux images présentant la plus grande dissimilarité, assurant ainsi une couverture maximale du contenu de la vidéo.
- Par la suite, l'algorithme ajoute itérativement la nouvelle image qui maximise la distance minimale par rapport à l'ensemble des images déjà sélectionnées. Cette stratégie permet de réduire les redondances tout en garantissant une représentation exhaustive des variations visuelles de la séquence.

5.4 Résultats et performances

- La méthode a été évaluée sur une base de vidéos consommateurs et comparée aux images clés établies par des juges humains.

- Les résultats montrent une correspondance satisfaisante avec la vérité terrain et soulignent la robustesse de l'approche face aux variations de perspective et de luminosité caractéristiques des vidéos non structurées.
- L'utilisation de l'image epitome combinée à la divergence de Kullback-Leibler permet de réduire significativement le coût de calcul, rendant l'algorithme efficace même pour des volumes importants de données vidéo.

Figure 7: "Valeurs de vérités" de la vidéo

Figure 8: Image clés sélectionné par la méthode

5.5 Conclusions et perspectives

- L'approche présentée permet d'extraire efficacement des images clés sans recourir à des techniques de segmentation ou de détection de plans, grâce à l'utilisation intelligente de l'image epitome.
- Les expérimentations démontrent la pertinence de cette méthode, qui offre un compromis intéressant entre précision et efficacité computationnelle dans un contexte de vidéos consommateurs.

6 Méthode d'extraction d'images clés guidée par la qualité et la détection d'objets

Ce document [5] présente une approche pour extraire des images clés à partir d'un flux vidéo, en s'appuyant sur la détection d'objets et l'évaluation de la qualité des images. La méthode vise à réduire la redondance tout en conservant les informations pertinentes, grâce à une combinaison de techniques de détection par réseaux de neurones convolutionnels (CNN) et de mesures basées sur l'entropie.

6.1 Première étape : Détection d'objets et attribution d'un score de qualité

- Un détecteur d'objets est appliqué à chaque image pour identifier des éléments d'intérêt (par exemple, piétons, véhicules).
- Chaque image se voit attribuer un score de qualité en fonction de la présence et de la pertinence des objets détectés.
- L'utilisation de ce score permet de prioriser les images contenant des informations significatives et d'éliminer une grande partie des images redondantes.

6.2 Deuxième étape : Extraction de caractéristiques profondes par CNN

- Les images candidates sont ensuite traitées par un réseau de neurones convolutionnel, ici basé sur l'architecture AlexNet, pour extraire des représentations profondes.
- Les caractéristiques extraites par le CNN fournissent une description riche et sémantique des images, facilitant ainsi la distinction entre images informatives et images redondantes.
- Cette étape complète l'évaluation initiale basée sur le score de qualité en affinant la sélection des images clés.

6.3 Troisième étape : Sélection finale des images clés

- Les informations issues de la détection d'objets et des caractéristiques CNN sont fusionnées avec des mesures d'information mutuelle (basées sur l'entropie) et des descripteurs locaux (par exemple, SURF).
- Ces mesures permettent de quantifier la similarité entre les images et d'identifier celles qui représentent le mieux le contenu du flux vidéo.
- La méthode ajuste également la sélection en fonction de paramètres d'échantillonnage, assurant ainsi la continuité temporelle et la pertinence des images retenues.

6.4 Résultats et performances

- Des expérimentations sur des vidéos de surveillance (notamment dans le contexte ferroviaire) démontrent que l'approche permet d'extraire des images clés plus précises et moins redondantes.
- La méthode présente une amélioration notable par rapport aux techniques traditionnelles basées sur des caractéristiques SIFT ou des méthodes de clustering, tant en termes de précision que d'efficacité de calcul.
- L'évaluation expérimentale confirme que l'intégration des scores de qualité et des caractéristiques profondes aboutit à une représentation fidèle et compacte du contenu vidéo.

6.5 Conclusions et perspectives

- L'approche proposée combine de manière efficace la détection d'objets, l'analyse de qualité d'image et l'extraction de caractéristiques profondes pour l'extraction d'images clés.
- En s'appuyant sur le potentiel des CNN et sur des mesures d'information basées sur l'entropie, la méthode parvient à sélectionner des images qui représentent fidèlement le contenu du flux vidéo, tout en réduisant la redondance.
- Les perspectives futures incluent l'optimisation des paramètres de sélection, l'adaptation de la méthode à d'autres types de vidéos, et l'intégration d'architectures de réseaux plus récentes pour améliorer encore la performance du système.

7 Méthode d'extraction d'images clés basée sur l'algorithme ISPMDE-SVM

Ce document [6] présente une approche d'extraction d'images clés à partir de vidéos, qui combine l'optimisation des paramètres d'un classifieur SVM avec une version améliorée de l'algorithme d'évolution différentielle. L'algorithme proposé, dénommé ISPMDE-SVM, vise à pallier les difficultés rencontrées dans la sélection manuelle ou arbitraire des paramètres SVM, afin d'améliorer la précision de l'extraction d'images clés.

7.1 Première étape : Extraction et prétraitement des caractéristiques vidéo

- Les caractéristiques globales du mouvement sont extraites à l'aide de l'algorithme optical flow, qui permet de quantifier les déplacements globaux entre images consécutives. Parallèlement, des caractéristiques locales sont obtenues via des méthodes de traitement d'image, notamment la conversion en espace HSV et le calcul d'histogrammes de couleur pour capturer les variations locales.

- À partir des mesures de mouvement (moyenne de distances, volatilité des changements locaux et globaux), un vecteur de caractéristiques est construit pour chaque séquence ou image. Ce vecteur résume les informations essentielles sur le contenu et le changement visuel de la vidéo.

7.2 Deuxième étape : Optimisation des paramètres SVM par l'algorithme ISPMDE

- Les méthodes traditionnelles de SVM rencontrent des difficultés pour déterminer automatiquement les paramètres optimaux, ce qui peut impacter négativement la précision de l'extraction.

- **Optimisation par évolution différentielle améliorée :** Pour remédier à ce problème, l'approche introduit un algorithme d'évolution différentielle dit "ISPMDE" (Improved Self Perturbation Mutation Differential Evolution). Cet algorithme combine des stratégies de mutation classiques (DE/rand/1/bin et DE/best/2/bin) en y intégrant une méthode de mutation à perturbation indépendante. L'objectif est de maintenir une diversité élevée dans la population tout en accélérant la convergence vers la solution optimale.

7.3 Troisième étape : Extraction des images clés via le modèle ISPMDE-SVM

- **Apprentissage et détection :** Les vecteurs de caractéristiques extraits des vidéos servent à constituer un jeu de données d'entraînement pour le SVM, dont les paramètres sont optimisés par l'algorithme ISPMDE. Une fois le modèle SVM entraîné, il est utilisé pour analyser de nouvelles séquences vidéo et détecter automatiquement les images clés qui représentent au mieux le contenu de chaque séquence.

- L'algorithme permet non seulement d'extraire des images clés de manière plus précise, mais aussi de réduire la redondance en ne sélectionnant que les images les plus représentatives pour la recherche ou l'indexation vidéo.

7.4 Résultats et performances

- Des tests effectués sur des fonctions standards et des bases de données vidéo (dont la TREC2009 et une base de tests interne) démontrent que l'algorithme ISPMDE-SVM offre une meilleure précision et un meilleur taux de rappel dans l'extraction des images clés, par rapport à des méthodes utilisant directement le SVM ou d'autres variantes de l'évolution différentielle.

- L'algorithme proposé atteint une convergence plus rapide tout en préservant la diversité de la population, ce qui se traduit par une optimisation efficace des paramètres SVM sans accroître significativement la complexité temporelle globale du processus.

7.5 Conclusions et perspectives

- L'approche ISPMDE-SVM permet d'extraire des images clés de manière autonome et précise en combinant une extraction fine des caractéristiques vidéo avec une optimisation intelligente des paramètres du classifieur SVM.

- Les résultats expérimentaux confirment l'efficacité de cette méthode, qui offre un bon compromis entre précision d'extraction et coût computationnel.

- Des travaux futurs pourraient explorer l'intégration d'autres types de caractéristiques vidéo ou l'adaptation de l'algorithme à des environnements de traitement en temps réel.

8 Méthode d'extraction d'images clés basée sur la transformée de Fourier quaternionique avec fusion de caractéristiques multiples

La méthode proposée dans ce document [7] vise à extraire des images clés (key frames) à partir de vidéos de surveillance en préservant l'intégralité de l'information visuelle et en capturant à la fois les mouvements globaux et locaux du ou des sujets.

8.1 Etape 1 : Extraction des caractéristiques

- **Caractéristiques dynamiques** : On utilise la différence entre deux images consécutives pour obtenir la carte du mouvement, notée FM(x, y), qui représente l'évolution temporelle de la scène.

- **Caractéristiques statiques** : Contrairement aux méthodes classiques qui convertissent en niveaux de gris, cette approche conserve les informations de couleur. On extrait ainsi le signal de luminosité FB(x, y) et deux caractéristiques de couleur inspirées du système de double opposition (rouge/vert et bleu/jaune), notées FRG(x, y) et FBY(x, y).

8.2 Etape 2 : Fusion par représentation quaternion

Les quatre caractéristiques sont ensuite fusionnées dans une unique représentation sous forme de quaternion, c'est-à-dire :

$$f(x,y) = FM(x,y) + FB(x,y)\mu_1 + FRG(x,y)\mu_2 + FBY(x,y)\mu_3$$

Cette représentation permet d'incorporer simultanément les informations de mouvement et de couleur dans un espace multidimensionnel.

8.3 Etape 3 : Transformation de Fourier en domaine quaternion

- Une transformation de Fourier dans le domaine quaternion est appliquée sur la représentation fusionnée.
- L'intérêt majeur ici est de récupérer le spectre de phase de l'image, lequel contient des informations essentielles sur les contours et la structure globale de la scène, contrairement au spectre d'amplitude qui reflète surtout la distribution de niveaux de gris.

8.4 Etape 4 : Filtrage et reconstruction

- Le spectre de phase obtenu est ensuite filtré par un filtre gaussien afin de réduire le bruit et les composants de haute fréquence, ce qui permet de mieux isoler la structure de l'image.
- Une transformation de Fourier inverse est alors appliquée pour reconstruire une carte de caractéristiques fusionnées (fused feature map) qui met en évidence les contours et détails importants de la scène.

8.5 Etape 5 : Sélection adaptative des key frames

- Pour identifier les moments significatifs, on calcule la différence (via l'erreur quadratique moyenne, ou MSE) entre les cartes de caractéristiques de deux images consécutives.
- En construisant une courbe de ces différences, les points extrêmes (indiquant des changements marqués dans l'état de mouvement) sont identifiés comme candidats pour être des images clés.
- Des seuils adaptatifs, définis à partir des différences moyennes entre images voisines, permettent ensuite de filtrer ces candidats pour ne retenir que les images qui représentent réellement un changement notable dans le contenu vidéo.

8.6 Résultats

8.6.1 Précision de l'extraction

La méthode parvient à extraire avec justesse les images clés en capturant les changements dans les états de mouvement global et local des cibles. Par exemple, dans une vidéo illustrant une personne qui entre dans le champ de la caméra, marche, s'agenouille et se relève, les images clés sélectionnées reflètent de manière précise chacune de ces phases de l'action.

8.6.2 Comparaison avec d'autres approches

Lors de comparaisons avec d'autres méthodes basées sur le domaine fréquentiel (FD, FFT) et sur l'analyse d'échelle et de direction (SaD), le modèle proposé montre des performances supérieures.

8.6.3 Analyse

Les tests sur des vidéos complexes, comme celles comportant plusieurs cibles et des événements d'anomalie (vol, agression, vandalisme), illustrent que la méthode capture non seulement les changements globaux (entrée et sortie de la scène) mais également les détails locaux significatifs (mouvements de mains, changement de posture) avec une grande fidélité.

8.7 Conclusion

En résumé, la méthode se distingue par sa capacité à :

- Conserver l'intégralité de l'information de couleur (en évitant le prétraitement en niveaux de gris)
- Fusionner intelligemment des caractéristiques dynamiques et statiques via une représentation quaternion
- Exploiter le spectre de phase pour extraire avec précision les contours et mouvements
- Appliquer une sélection adaptative basée sur des mesures de changement d'image (MSE) pour extraire les images clés les plus pertinentes

Cette approche permet ainsi d'améliorer la précision de l'extraction des key frames en conservant à la fois les détails globaux et locaux dans des vidéos de surveillance.

9 Références

1 - Wisnu Widiarto & Eko Mulyanto Yuniarno & Mochamad Hariadi, "Video Summarization Using a Key Frame Selection Based on Shot Segmentation", 2015 International Conference on Science in Information Technology (ICSITech).

2 - Yuli Gao & Tong Zhang & Jun Xiao, "THEMATIC VIDEO THUMBNAIL SELECTION", 2009 16th IEEE International Conference on Image Processing (ICIP).

3 - Mrityunjay Kumar & Alexander C. Loui, "KEY FRAME EXTRACTION FROM CONSUMER VIDEOS USING SPARSE REPRESENTATION", 2011 18th IEEE International Conference on Image Processing.

4 - C.T.Dang & M.Kumar & H.Radha, "KEY FRAME EXTRACTION FROM CONSUMER VIDEOS USING EPITOME", 2012 19th IEEE International Conference on Image Processing.

5 - Mingju Chen & Xiaofeng Han & Hua Zhang & Guojun Lin & M.M. Kamruzzaman, "Quality-guided key frames selection from video stream based on object detection", 2019 J. Vis. Commun. Image R.

6 - Xiao-Gen PEI, "The key frame extraction algorithm based on the indigenous disturbance variation difference video", 10th International Conference of Inforamtion and Communication Technology (ICICT-2020).

7 - Yunzuo Zhang & Jiayu Zhang & Ruixue Liu & Pengfei Zhu & Yameng Liu, "Key frame extraction based on quaternion Fourier transform with multiple features fusion", Expert Systems With Applications Volume 216 (2023).