Stage M2

Synthèse Méthode Extraction de *Tag Image*

Benjamin Serva
Master 2 IMAGINE
Université de Montpellier

15 mai 2025

Encadrants :

Olivier Strauss & William Puech & Frédéric Comby

1 Introduction

Ce document a pour objectif de synthétiser l'implémentation et les tests réalisés sur trois méthodes d'extraction automatique de *tag image* à partir de vidéos. Il présente également les points positifs et négatifs de chaque méthode, ainsi que leurs perspectives d'amélioration.

2 Méthode 5 : basée sur la détection d'objets

2.1 Explication

Cette méthode combine à 50/50 un score SIFT et un score basé sur la détection d'objets, en prenant en compte le taux de confiance ainsi que la taille des objets détectés.

Plusieurs paramètres sont ajustables :
-Le taux de confiance minimal pour qu'un objet soit pris en compte (supérieur à 0,7)
-La taille minimale (1% de l'image) et maximale (40% de l'image) des objets considérés
-Une puissance (valeur 3) utilisée pour accentuer l'influence de la taille des objets dans le calcul du score
-Un bonus spécifique attribué à la détection d'un humain (score de 10)

2.2 Avantages / Inconvénients

- **Avantages :**
  - Évite de sélectionner une frame vide ou non informative. Même si la frame choisie n'est pas optimale, elle contient toujours un minimum d'informations, ce qui rend cette méthode plus stable que d'autres.
  - Méthode très modulable et paramétrable par l'utilisateur.
- **Inconvénients :**
  - A tendance à privilégier les frames contenant de nombreux objets.
  - Favorise fortement les frames dans lesquelles des humains sont détectés.

2.3 Perspectives d'amélioration

1. Appliquer une fonction de pondération temporelle qui favorise les frames situées entre le milieu et la fin du plan.
2. Ajouter un score basé sur la position des objets dans l'image : un objet situé au centre aurait un poids plus important qu'un objet en bord d'image.
3. Mettre en place un critère de qualité d'image pour éviter la sélection de frames floues.
4. Favoriser les moments où la caméra est immobile plutôt que ceux où elle est en mouvement.
5. Étant donné qu'on peut gérer le nombre de *tag image*, il peut être intéressant d'en extraire plusieurs, puis d'en sélectionner une en fonction de critères secondaires.

3 Méthode 6 basé sur le mouvement

3.1 Explication

Chaque frame est transformée en un vecteur de caractéristiques basé sur la couleur (HSV) et le mouvement (flux optique). À l'aide d'une fenêtre glissante, on extrait des statistiques (changement de texture et intensité du mouvement) pour former des vecteurs $X[i]$.

Un One-Class SVM avec noyau RBF est ensuite entraîné pour détecter les frames atypiques, avec une optimisation des hyperparamètres (ν, γ) via Differential Evolution (DE). Enfin, les frames ayant les scores les plus négatifs sont sélectionnées comme *tag image*.

3.2 Avantages / Inconvénients

- **Avantages :**
  - Sélectionne une frame dans laquelle il y a du mouvement.
- **Inconvénients :**
  - Dans certains cas, la frame présentant le plus de mouvement ne correspond pas au moment contenant le plus d'informations utiles.
  - Il arrive que la frame sélectionnée soit floue.

3.3 Perspectives d'amélioration

1. Mettre en place un critère de qualité d'image pour éviter la sélection de frames floues.
2. Plutôt que de chercher la frame la plus atypique, on peut choisir de faire l'inverse en sélectionnant la frame la plus courante.

4 Méthode 7 basé sur la fusion de caractéristiques en quaternion

4.1 Explication

- Pour chaque image, on extrait quatre canaux de caractéristiques : FM (motion) : différence absolue entre l'image courante et la précédente (si elle existe), luminance, FRG (red/-green) et FBY (blue/yellow).
- Fusion par transformée de Fourier quaternion : on regroupe ces quatre caractéristiques en deux plans complexes, Plan 1 : FM + i·FB et Plan 2 : FRG + i·FBY. On effectue une FFT 2D sur chacun des deux plans puis on calcule une carte de phase fusionnée. On normalise cette phase en image 8 bits (0–255).
- Filtrage spatial : On applique un flou gaussien sur la carte de phase normalisée, afin d'atténuer le bruit (les hautes fréquences).
- Reconstruction (inverse FFT): On traite la carte floutée comme un signal réel et on fait une IFFT 2D. Le module du résultat constitue la carte fusionnée finale pour cette frame, qui met en valeur à la fois le mouvement et les variations de couleur/brillance.

- Détection de la frame la plus saillante: On calcule la MSE (Mean Squared Error) entre chaque carte fusionnée et celle de la frame précédente, pour obtenir une courbe MSE tout au long de la vidéo. L'indice où cette MSE est maximale correspond à la transformation la plus marquée (changement global + local).

4.2 Avantages / Inconvénients

- **Avantages :**
  - Beaucoup plus rapide que les deux autres méthodes.
- **Inconvénients :**
  - Très dépendante de l'intensité du flou appliqué, qui peut varier d'une vidéo à l'autre.

4.3 Perspectives d'amélioration

Aucune amélioration réellement envisageable. La méthode repose sur un principe spécifique et complexe (dont l'utilisation n'est pas forcément justifié), et ses résultats sont très variables et globalement peu convaincants.

5 Possibilités à explorer pour plus tard

5.1 Plans avec caméra fixe

Dans ce cas, il serait pertinent de détecter les éléments présents en permanence dans la scène (car ils ne bougent pas et sont donc peu informatifs), afin d'isoler ceux qui apparaissent temporairement.

5.2 Analyse sémantique avec YOLOv8

Une première passe par détection d'objets (via YOLOv8) permettrait de connaître le nombre d'occurrences de chaque objet ainsi que leur pourcentage d'apparition dans la vidéo. Cette information pourrait être utilisée pour adapter la stratégie d'extraction de la *tag image* : par exemple, utiliser un algorithme spécifique lorsqu'un humain est détecté, et un autre dans le cas contraire.

5.3 Détection de faibles variations dans la vidéo

La méthode développée repose sur le SSIM, qui mesure la similarité entre chaque paire de frames successives. Si la moyenne des scores SSIM dépasse un certain seuil, on considère que le contenu visuel varie peu. Plusieurs améliorations sont possibles, notamment :

- L'utilisation d'une fenêtre glissante pour affiner la détection des plans à faibles variations
- Des comparaisons SSIM entre des images moyennes (moyenne temporelle d'un plan) et les images individuelles de la vidéo.