POLYTECHNIQUE MONTRÉAL

UNIVERSITÉ D'INGÉNIERIE

INF8870

Introduction aux Technologies Multimédias
Rapport

TP3 - Indexation de contenu pictural

Élèves :
Marc Zhang 2312403
Clément Auclin 2308904

Enseignant :
Wissal Zarrami

14 novembre 2024

Table des matières

1 Introduction

Ce TP a pour objectif d'implémenter un algorithme permettant de détecter l'appartenance d'une image à une vidéo. Pour cela, deux types d'images différentes seront utilisés : en JPEG et en PNG. Les vidéos sont quand à elles en .mp4. Au total, nous disposons de 23min23s de vidéos, décomposées en 100 différentes, toutes entre 5 et 25 secondes.

2 Question 1

2.1 Implémentation d'Alice

La méthode d'Alice possède de nombreux problèmes :

Le premier est qu'en dehors même de la réussite ou non des opérations, l'implémentation sera particulièrement lente. En effet, le fait de comparer chaque frame de chaque vidéo à l'image recherchée vient alourdir considérablement les calculs. En considérant que les vidéos soient toutes en 30 fps (ce qui n'est pas sur, mais permet une bonne approximation ici), il faudra comparer l'image à 42 090 frames en tout.

Ensuite, l'utilisation d'une égalité pour retourner la vidéo de laquelle est tirée l'image n'est pas une bonne idée : la compression des fichiers JPEG ne permettra jamais d'obtenir une différence nulle avec les frames de la vidéo à cause de la compression par rapport aux PNG originaux. Ainsi, toutes les images JPEG seront considérées comme ne faisant pas partie d'une vidéo, d'autant plus que c'est une différence euclidienne directement entre chaque pixel de l'image qui est utilisée (ce qui est alors très sensible).

2.2 Implémentation de Bob

Bob reprend certains problèmes d'Alice :

Il souhaite lui aussi traverser image par image, ce qui est très mauvais d'un point de vue du temps de calcul. De plus, il veut conserver chaque image à part dans un fichier. Cela n'a aucun intérêt, les vidéos étant déjà composées d'images successives.

Cependant, s'il souhaite réellement implémenter cette méthode, il doit utiliser le format PNG qui ne crée pas de pertes dans les informations des images. En effet, la compression avec perte JPEG, viendra rendre plus difficile les traitements et les comparaisons même avec les autres images JPEG, car on ne dispose pas des paramètres particuliers utilisés pour compresser les images (et donc ceux-ci pourraient être différents des paramètres pour la compression des vidéos).

2.3 Implémentation de Carol

La méthode de Carol permet de réparer certaines erreurs commises par Alice et Bob :

La sélection d'une image par seconde permet d'alléger considérablement les calculs. De plus, l'utilisation d'histogrammes permet d'être moins sensible au bruit et donc d'obtenir possiblement de meilleurs résultats que ce soit en PNG ou en JPEG. Aussi, Carol corrige le problème de la différence nulle des deux méthodes précédentes en récupérant cette fois

la vidéo dont la distance est minimum : cette méthode fonctionne donc techniquement à la fois pour PNG et pour JPEG.

Cependant, cette méthode n'utilisant pas de seuil, les images ne seront jamais considérées comme n'appartenant à aucune vidéo et donc l'ensemble des images devant être "out" ne le seront pas, engendrant des erreurs.

3 Question 2

Il a été choisi d'implémenter l'algorithme de Carol. En effet, comme expliqué précédemment, il s'agit du plus apte à fonctionner. Il pourra alors servir de socle et être adapté pour la question suivante.

Pour les calculs des histogrammes, la concaténation est effectuée dans l'ordre bleu, vert puis rouge, ce qui donne un vecteur de 256*3 = 768 éléments pour décrire chaque image.

Pour la détection des histogrammes dans les vidéos, comme prévu par Carol, une image est récupérée toutes les secondes. Cependant, cette récupération provient d'un arrondi qui cherche à obtenir l'image la plus proche (certaines vidéos ont un nombre de frames par secondes qui n'est pas exactement un entier). On conserve alors dans le dictionnaire des histogrammes le temps exact de la frame en multipliant l'indice de l'image par le nombre de fps.

Les résultats obtenus sont les suivants :

	Images PNG	Images JPEG
Taux de bonnes réponses	78.0%	70.3%
Nombre de bonnes prédictions (sur 300)	234	211
Écart temporel moyen (en s)	0.47	2.73

Table 1 – Résultats de la méthode de Carol

Les résultats obtenus sont ainsi bien cohérents avec les hypothèses formulées pour la question 1 : l'algorithme est moins performant sur les images JPEG (bien qu'il soit tout de même capable de créer des résultats corrects) du à leur compression mais parvient cependant à générer un écart moyen en secondes relativement faible grâce à la récupération d'une image toute les secondes.

Le problème majeur réside donc dans le fait qu'aucune image ne peut être catégorisée comme "out" du à l'absence de seuil.

Par ailleurs, les différents temps d'exécution permettent d'attester que cette méthode est plutôt efficace :

	Temps (en s)	Temps par element
Traitement des vidéos	107	0.93 vidéos/s
Traitement des PNG	15	20 images/s
Traitement des JPEG	11	27.27 images/s

Table 2 – Temps d'exécution des différentes parties de l'algorithme

4 Question 3

Cette partie permet de répondre à la question 3, en explicitant la démarche réalisée. Quatre techniques ont été essayées :

- Calcul des histogrammes en rajoutant un seuil
- Utilisation des HOG (Histogram of Oriented Gradients) en remplacement des histogrammes, toujours avec un seuil
- L'utilisation d'un descripteur fourni par l'avant dernière couche d'un modèle ResNet pré-entrainé (en y ajoutant un seuil)
- La combinaison de la partie ResNet pour la détection de la vidéo et de la partie HOG pour la détection de l'instant à l'intérieur de celle-ci

C'est finalement la quatrième méthode qui a été conservée.

4.1 Utilisation d'histogrammes et de seuils

Il a dans un premier temps été essayé d'améliorer le score de la méthode de Carol en ajoutant un seuil pour pouvoir affirmer si une image n'est dans aucune vidéo (ce qui n'était pas possible précédemment). Le nombre d'images récupérées par vidéo n'a lui pas changé afin de permettre d'avoir un écart de temps le plus faible possible pour les résultats, tout en tenant compte que chaque vidéo n'est composée que d'une unique prise de vue (cette fréquence d'échantillonage sera conservée pour l'ensemble des méthodes suivantes).

Ainsi, nous pouvons obtenir le pourcentage de bonnes réponses et l'écart temporel en fonction du seuil :

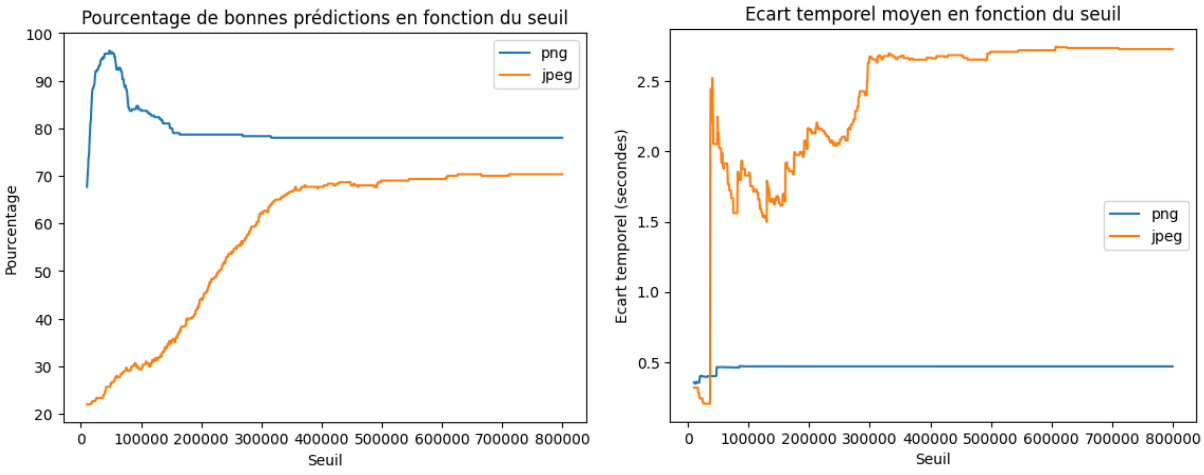

FIGURE 1 – Pourcentage de bonne réponses et écart temporel en fonction du seuil avec la méthode des histogrammes

Nous obtenons d'excellents résultats pour le PNG (jusqu'à 95 %) en utilisant le seuil approprié. Cependant, le seuil qui maximise la recherche PNG se situe dans une plage où la recherche JPEG est très mauvaise (environ 30 %). Étant donné que nous devons mettre en œuvre un algorithme de recherche qui ne tient pas compte du format de l'image, nous sommes obligés de choisir une valeur de seuil très élevée, ce qui rend l'utilisation d'un seuil peu intéressante.

4.2 Utilisation des HOG

Notre deuxième approche a consisté à ne pas comparer les histogrammes des images, mais plutôt les HOG. Dans un premier temps, nous avons implémenté la recherche sans seuil. Voici les résultats obtenus :

	Images PNG	Images JPEG
Taux de bonnes réponses	76.0%	64.3%
Nombre de bonnes prédictions (sur 300)	228	192
Écart temporel moyen (en s)	0.31	0.52

TABLE 3 – Résultats de la méthode des HOG

On remarque que les résultats sont moins satisfaisants au niveau de la prédiction par rapport à la méthode utilisant les histogrammes. Cependant, il est à noter que l'écart temporel moyen est très faible.

Dans un deuxième temps, nous avons implémenté la recherche à partir des HOG avec l'introduction d'un seuil. Voici les résultats obtenus :

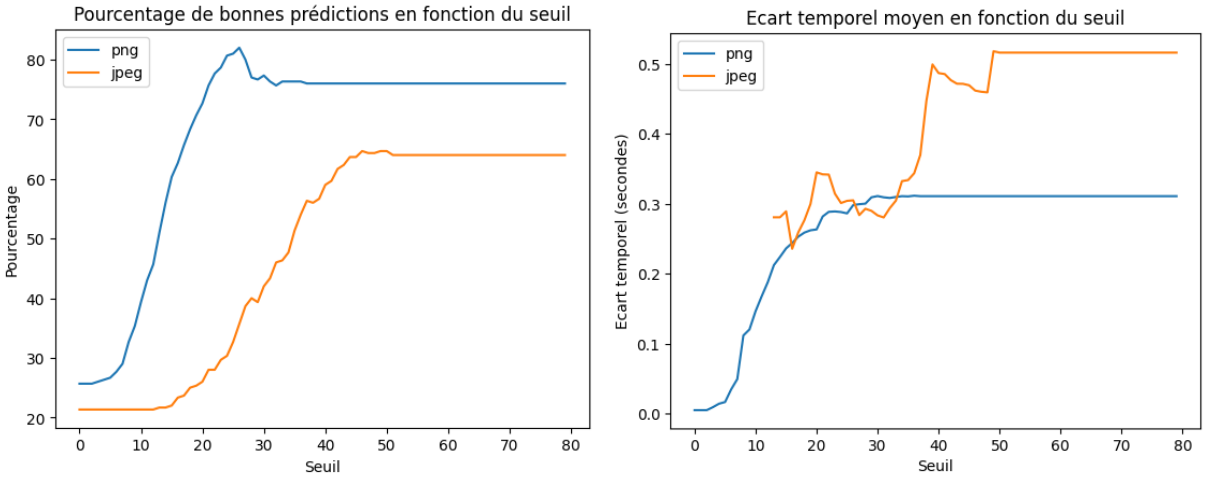

FIGURE 2 – Pourcentage de bonne réponses et écart temporel en fonction du seuil avec la méthode des HOG

Nous observons un pourcentage de bonnes prédictions plus faible que celui obtenu avec la méthode des histogrammes, indépendamment du seuil choisi, ce qui rend cet algorithme inutile. Cependant, l'écart temporel moyen est lui très satisfaisant.

4.3 Utilisation d'un modéle ResNet

Notre troisième approche a consisté à utiliser un réseau neuronal résiduel pour décrire nos images. Ici, ce sera le modèle ResNet50 fourni par la librairie Keras. À partir de ce modèle, nous avons comparé les vecteurs obtenus à partir de l'avant-dernière couche du réseau neuronal, qui, bien qu'ils soient illisibles, constituent un descripteur intéressant pour la comparaison des caractéristiques des images.

Voici les résultats de l'implémentation sans l'utilisation de seuils :

	Images PNG	Images JPEG
Taux de bonnes réponses	78.0%	77.3%
Nombre de bonnes prédictions (sur 300)	234	232
Écart temporel moyen (en s)	2.46	2.91

TABLE 4 – Résultats de la méthode du modèle ResNet sans seuil

Voici les résultats de l'implémentation avec l'utilisation de seuils :

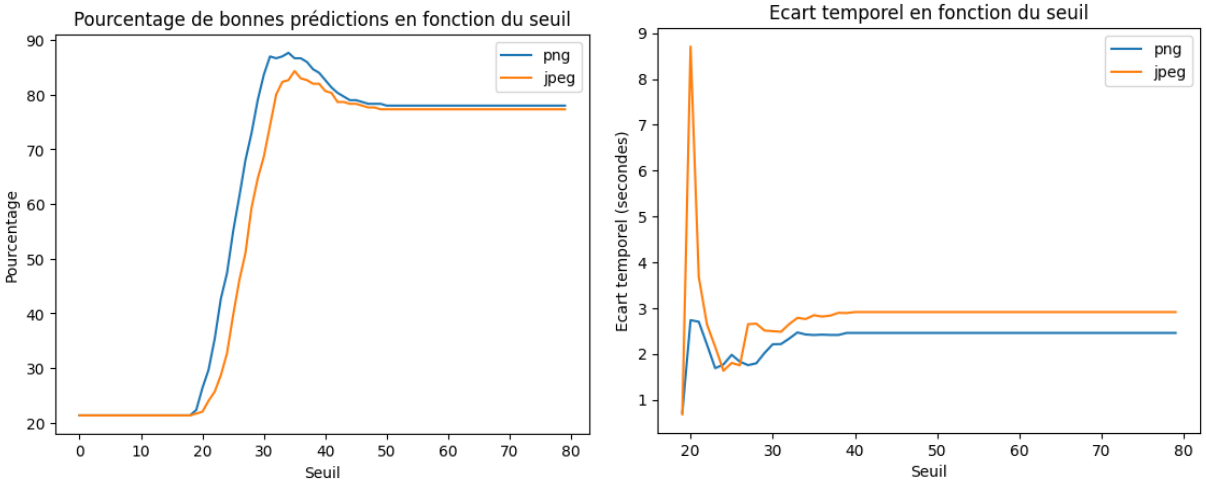

FIGURE 3 – Pourcentage de bonne réponses et écart temporel en fonction du seuil avec la méthode du modèle ResNet

Nous obtenons de très bons résultats. Bien qu'ils ne soient pas aussi bons que ceux obtenus avec l'utilisation d'histogrammes sur les fichiers PNG avec le seuil optimal, nous avons l'avantage ici d'atteindre un pourcentage maximal de bonnes prédictions pour le format JPEG et PNG pour la même valeur de seuil. Cela en fait une méthode de recherche d'image intéressante. Cependant, l'inconvénient de cette approche est la présence d'un écart temporel plutôt élevé en comparaison avec les deux précédents algorithmes.

4.4 Méthode finale : combinaison du modèle ResNet avec l'utilisation des HOG

Pour obtenir la meilleure méthode de recherche, nous avons finalement choisi de combiner l'utilisation du modèle ResNet avec l'utilisation des HOG pour les comparaisons. Dans un premier temps, l'algorithme recherche une vidéo correspondant à l'image à l'aide du modèle ResNet. Ensuite, si une vidéo est trouvée, l'algorithme recherche le minutage de la vidéo dans lequel l'image apparaît en comparant les HOG. Voici les résultats obtenus :

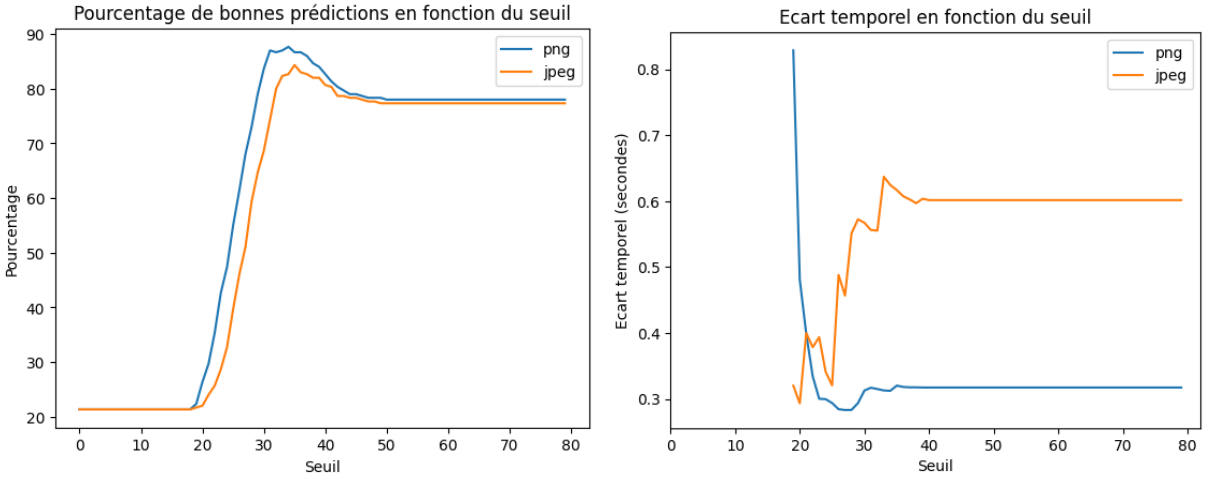

FIGURE 4 – Pourcentage de bonne réponses et écart temporel en fonction du seuil avec la combinaison des méthodes

Finalement, nous obtenons de très bons résultats tant pour le format JPEG que PNG, que ce soit en pourcentage de bonnes prédictions ou en écart temporel.

Le seuil final fixé est de 35. Pour cette valeur, voici nos résultats :

	Images PNG	Images JPEG
Taux de bonnes réponses	86.7%	84.3%
Nombre de bonnes prédictions (sur 300)	260	253
Écart temporel moyen (en s)	0.32	0.62

TABLE 5 – Résultats de la combinaison des méthodes

De plus, voici les différents temps de calculs des étapes de l'algorithme :

	Temps (en s)	Temps par element
Traitement des vidéos	222	0.45 vidéos/s
Traitement des PNG	50.5	5.94 images/s
Traitement des JPEG	43.9	6.83 images/s

TABLE 6 – Temps d'exécution des différentes parties de l'algorithme final

Le traitement des vidéos est bien effectué en moins de 5 minutes sur l'ordinateur portable utilisé.

Processeur : AMD Ryzen 5 5600H with Radeon Graphics 3.30 GHz

RAM : DDR4 16Go

Carte graphique : NVIDIA GeForce RTX 3060 Laptop GPU

4.5 Comparaison des performances avec l'algorithme de Carol

Pour conclure, il est intéressant de venir comparer les résultats de cet algorithme avec celui de Carol. Les informations les plus importantes sont récapitulées dans le tableau suivant :

	Méthode de Carol	Méthode personnelle
Traitement des vidéos (en s)	107	222
Taux de bonnes réponses PNG	78.0%	86.7%
Taux de bonnes réponses JPEG	70.3%	84.3%
Écart temporel PNG (en s)	0.47	0.32
Écart temporel JPEG (en s)	2.73	0.62

TABLE 7 – Comparaison des deux méthodes implémentées

On remarque alors que la méthode implémentée dans la question 3 est bien plus performante que celle de la question 2. En effet, l'ensemble des mesures de performances sont meilleures (que ce soit les taux de bonnes réponses ou les écarts temporels) à la fois pour les PNG et les JPEG. Cependant, comme il était possible de s'y attendre, la méthode de la question 3 prend plus de deux fois le temps de celle de la question 2 pour traiter les vidéos. Cela s'explique par l'utilisation des gradients et d'un modèle d'IA (ResNet50) qui sont bien plus gourmands en calculs que la méthode des histogrammes de Carol.

Au final, bien que notre méthode nous donne de très bons résultats au niveau de l'écart temporel, elle reste limitée au niveau du pourcentage de bonne prédiction qui, bien que bon, n'est pas parfait (ne parvenant pas à atteindre les 90% et étant ainsi inférieur pour les PNG à la méthode des histogrammes de couleurs).

5 Question 4

L'algorithme utilisé ici est donc le même que précédemment. Il est seulement possible de récapituler les temps de calculs pour les différentes étapes, puisque les fichiers de validation nous sont évidemment indisponibles :

	Temps (en s)	Temps par element
Traitement des vidéos	224	0.45 vidéos/s
Traitement des PNG	54.1	5.54 images/s
Traitement des JPEG	43.7	6.86 images/s

TABLE 8 – Temps d'exécution des différentes parties de l'algorithme final