Détection de contenu explicite dans les vidéos

Hugo Jean, Emmanuel Giguet, Christophe Rosenberger

Détection de contenu explicite dans les vidéos

Hugo Jean Emmanuel Giguet Christophe Rosenberger

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

{hugo.jean, emmanuel.giguet, christophe.rosenberger}@unicaen.fr

Résumé

L'analyse de contenu vidéo joue une rôle majeur dans la protection des enfants sur Internet et dans le travail des forces de l'ordre lors d'investigation numérique. Avec la rapide croissance de la taille des supports de stockage et les plateformes de partage, le besoin de solutions rapides et robustes d'analyse de vidéos devient de plus en plus nécessaire. Dans cet article, nous proposons un modèle de détection de contenu explicite dans les vidéos. Notre approche se base l'utilisation de réseau convolutionnel pour extraire des paramètres visuels de haut niveau et un réseau à mémoire pour exploiter la dimension temporelle de la vidéo. Nous validons notre approche sur un jeu de données composé de vidéos (avec différentes résolutions et durées) démontrant son intérêt opérationnel.

Mots clefs

Analyse de vidéos, apprentissage profond, contenu explicite.

1 Introduction

De nos jours, la diffusion de vidéos représente 80% du trafic sur Internet. Les contenus vidéos sont transmis dans des volumes en croissance exponentielle, la taille des outils de stockage et la réduction de leur prix permettent un stockage quasi illimité de vidéos. Ainsi, un individu peut facilement stocker chez lui plusieurs milliers de vidéos.

Dans un contexte d'enquête criminelle, les supports de stockage de suspects sont en général perquisitionnés pour la recherche de preuves. Les crimes tels que l'exploitation d'enfants et le partage non autorisé de contenu à caractère sexuel ont de plus en plus comme support les vidéos. Cependant, la taille de ces données et leur nature ne permet pas un traitement rapide par les forces de l'ordre, qui doivent souvent exploiter ces données sous la pression d'un délai, typiquement la durée d'une garde a vue. Ces contraintes sont propices à des erreurs humaines (données compromettantes non détectées). Aujourd'hui, les avancées technologiques en matière de traitement automatique des images ne sont plus à présenter, leur utilisation est universelle. Ceci ouvre des opportunités pour le développement de nouvelles méthodes automatiques et robustes.

L'objectif de ce travail est de proposer une méthode de détection du caractère explicite d'une vidéo avec de meilleures performances par rapport à l'état de l'art, à la fois en considérant le taux de reconnaissance et le temps de calcul. L'analyse est globale mais peut être aussi appliquée pour identifier des frames avec du contenu explicite. Cette méthode vise à être utilisée de façon opérationnelle dans des investigations numériques de disques durs. D'autres usages peuvent être envisagés comme la diffusion de vidéos avec contrôle parental ou la vérification automatique du contenu de vidéos lors d'un téléchargement sur un serveur.

Dans cet article, nous proposons une approche basée sur l'utilisation de modèle convolutionnel pour l'extraction de paramètres visuels de haut niveau. Nous utilisons ensuite un réseau avec mémoire pour utiliser la dimension temporelle d'une vidéo. Notre architecture est présentée dans la Figure 2 avec une vue de haut niveau.

Le plan de l'article est le suivant. La section 2 décrit les principales méthodes de la littérature sur la détection du caractère explicite dans les vidéos. La méthode proposée est décrite dans la section 3. La section 4 présente le protocole expérimental et les résultats obtenus. Nous concluons cet article dans la section 5 et définissons plusieurs perspectives à ce travail.

2 Etat de l'art

Le papier [1] proposant le dataset utilisé dans cet article recense plusieurs travaux et donne leurs performances relatives dans le tableau 1. La méthode θ_x consiste à simplement compter le nombre d'images considérées comme explicites et considérer la vidéo comme explicite si cette valeur dépasse un seuil défini. L'approche θ_y réalise la même chose mais avec un pourcentage de la vidéo, par exemple si 10% de toutes les frames sont explicites alors la vidéo est considérée comme explicite elle aussi. Enfin, la méthode θ_z compte le nombre de frames **successives** explicites et ainsi considère la vidéo explicite si par exemple 5 frames successives sont explicites.

Les résultats présentés ici se basent tous sur une analyse séquentielle des frames, soit en analysant toutes les frames ou alors jusqu'à ce qu'un seuil soit atteint (seuil de frames

Modèle	θ_x Compteur d'image	θ_y % d'image	θ_z images successives explicite
Mask R-CNN	85.63%	86.38%	85.63%
YOLOv4	87.25%	87.75%	87.00%
SSD	81.16%	83.48%	82.88%
Cascade Mask R-CNN	84.88%	86.63%	86.13%

TABLEAU 1 – *Performance des méthodes de l'état de l'art sur la base LSPD.*

classifiées comme explicites par exemple). L'utilisation de toutes les frames d'une vidéo permet en effet de prendre une décision globale, cependant des réseaux convolutionnel 3D utilisant un groupe de frames et non plus une seule frame ont prouvé leur utilité et leur robustesse. On peut citer par exemple, X3D [2] ou encore Resnet3D [3]. Ces réseaux utilisent une sélection de frames tirées d'une méthode d'échantillonage définie, par exemple ci dessous l'architecture des réseaux X3D.

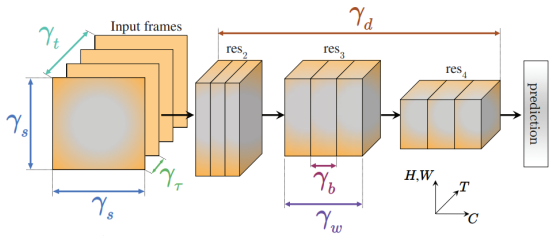

FIGURE 1 – *Base de X3D.*

Ici, nous nous intéresserons à γ_t tout simplement le nombre de frames en entrée du réseau et γ_τ l'écart entre les frames sélectionnées. Ces deux paramètres permettent entre autre d'influencer la taille du réseau et donc de choisir entre précision et rapidité. Le réseau X3D est pensé pour une utilisation sur mobile et est donc par définition rapide de base. Il existe cependant des versions plus complexes pour une utilisation classique.

3 Méthode proposée

L'approche proposée se base sur l'utilisation d'un modèle basé sur un réseau de neurones convolutionnel. Nous présentons tout d'abord le jeu de données utilisé pour l'entraînement du modèle et l'évaluation de performance. Nous détaillons ensuite le modèle ainsi que la stratégie d'analyse.

3.1 Dataset

Nous utilisons ici le dataset LSPD [1], celui ci contient 4000 vidéos (2000 de chaque classe) avec la répartition en durée dans la figure 2. Il est actuellement le dataset disponible au public le plus gros et le plus divers, ils est composé d'une partie image et vidéo, ici nous nous intéressons uniquement à la partie vidéo.

Label	< 1min	<5min	<10min	<20min	>20 min
Explicite	746	745	233	179	106
Normal	986	661	175	125	53

TABLEAU 2 – *Durée des vidéos dans LSPD.*

3.2 Modèle proposé

Le modèle proposé utilise un extracteur de paramètres de haut niveau ici X3D-M pour la vidéo, avec γ_t = 16 et en utilisant non pas une distribution uniforme comme proposé dans le papier original, mais une distribution normale décalée vers la fin de la vidéo. En effet, dans les vidéos, les contenus explicites sont souvent situés au milieu et à la fin. Nous utilisons ensuite un RNN et enfin un classifieur classique. Le modèle est décrit dans la figure 2. Cette version M de X3D fournit en sortie d'extracteur des vecteurs de taille 2048 que nous réduisons à une taille fixe de 512 par soucis de modularité.

Cette architecture permet une utilisation modulaire, on peut facilement ajouter une modalité (typiquement l'audio de la vidéo) et réaliser une fusion de paramètres avant le passage dans le RNN. Cela nous permettra dans le futur de pouvoir réaliser des tests sur l'intérêt et la robustesse de ces méthodes multimodales. Cela nous permet aussi d'inter-changer facilement et rapidement l'extracteur de paramètres, le classifieur et le réseau à mémoire.

4 Évaluation de la performance

Nous définissons dans cette section, le protocole et les résultats obtenus.

4.1 Protocole expérimental

Sur les 4000 vidéos composant le dataset, nous en sélectionnons 80% pour le set d'entraînement, 10% pour le set de validation et enfin 10% pour la base de test. Ainsi, nous avons 3200 vidéos (1600 par classe) d'entraînement, 400 de test et de validation.

L'implémentation de ce modèle est réalisé sous PyTorch, le modèle a été entraîné sur un cluster de 7 cartes Nvidia 1080 Ti avec 11 Go chacune. Cependant, il est important de noter que nous utilisons un cluster pour accélérer l'entraînement et non pas parce que nous sommes bridés par la taille du modèle (comme expliqué plus haut nous utilisons un modèle principalement utilisé pour les téléphones lors de l'inférence).

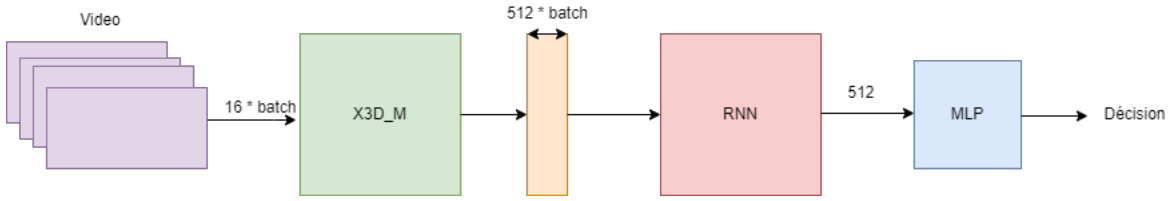

FIGURE 2 – *Architecture du modèle.*

4.2 Résultats

Le tableau 3 présente les résultats obtenus lors de l'évaluation de notre modèle sur notre base de test. Nous avons aussi proposé un modèle sans RNN et directement un classifieur après l'extraction des paramètres.

Head type	Performance
MLP	94%
RNN	96%

TABLEAU 3 – *Résultats sur LSPD.*

Les résultats obtenus sont excellents pour une classification binaire et sont meilleurs que les résultats par utilisation de modèle par frame unique présentés dans le tableau 1.

Lors de l'exécution du modèle pour la phase de test, nous avons calculé le temps d'inférence du modèle pour un paquet de 16 frames (voir le tableau 4). On peut noter un speedup d'environ 50 lors du passage sur GPU, ainsi l'utilisation d'un GPU est grandement favorable et reste assez facile d'accès de part la taille du modèle d'environ 16 MB en mémoire lors de l'inférence.

Processeur	Temps moyen d'inférence
CPU	907
GPU	20

TABLEAU 4 – *Temps d'inférence moyen (en ms).*

Nous présentons dans la figure 3 une illustration de l'analyse d'une vidéo en appliquant le modèle localement pour identifier les séquences avec un contenu explicite. Dans une investigation numérique opérationnelle, l'expert pourra définir un seuil et visualiser des résumés vidéos dans les séquences identifiées par la méthode proposée. Il pourra plus facilement décider si le contenu est répréhensible ou non.

5 Conclusion et perspectives

La méthode proposée permet une reconnaissance efficace et rapide du caractère explicite d'une vidéo facilitant grandement le travail d'investigation numérique sur un disque dur par exemple. L'approche proposée permet aussi d'identifier des séquences problématiques au regard de son contenu explicite. Des applications de l'approche peuvent également concerner le contrôle parental en masquant des frames inadéquates pour des enfants.

Dans cet article nous avons abordé la sélection de frames. Dans le papier où l'architecture X3D est proposé [2], la méthode tire uniformément γ_t frames avec un écart minimum de γ_τ. Pour notre entraînement, nous avons utilisé une distribution normale centrée sur la deuxième partie de la vidéo. Une des principales pistes de recherche est de pouvoir extraire une distribution réelle du contenu explicite dans les vidéos. On pourra par exemple utiliser les vidéos dont la durée est supérieure à 10 minutes et en analyser celles-ci frame par frame pour en tirer une distribution normalisée sur la durée de la vidéo. On pourra ensuite utiliser cette distribution en tant que méthode d'échantillonage de nos vidéos pour l'apprentissage. Cela devrait nous permettre d'accélérer l'apprentissage et d'améliorer la robustesse de nos modèles.

Un autre point déjà abordé dans ce papier est l'utilisation d'autre modalité pour notre classification par exemple l'audio qui semble être la modalité la plus facile d'accès. La classification d'audio aujourd'hui s'effectue à partir de spectrogramme de celui-ci, spectogramme ensuite fourni à un CNN pour bénéficier de l'état de l'art de ceux-ci comme par exemple proposé dans [4]. L'ajout de celle-ci pourrait se faire assez simplement due à la construction modulaire du réseau. Il suffirait d'utiliser une concaténation avant la couche RNN 4. On peut aussi mentionner les différentes techniques de fusion de paramètres disponibles pour les modèles multimodales proposées comme par exemple dans [5], ou même encore les méthodes par ensemble qui peuvent être utilisées pour obtenir de meilleurs résultats.

Références

[1] Phan Duy, Thanh Nguyen, Quang Nguyen, Hoang Tran, Ngoc-Khoi Khac, et Lung Vu. Lspd : A large-scale pornographic dataset for detection and classification. *International Journal of Intelligent Engineering and Systems*, 15 :198, 02 2022.

[2] Christoph Feichtenhofer. X3D : expanding architectures for efficient video recognition. *CoRR*, abs/2004.04730, 2020.

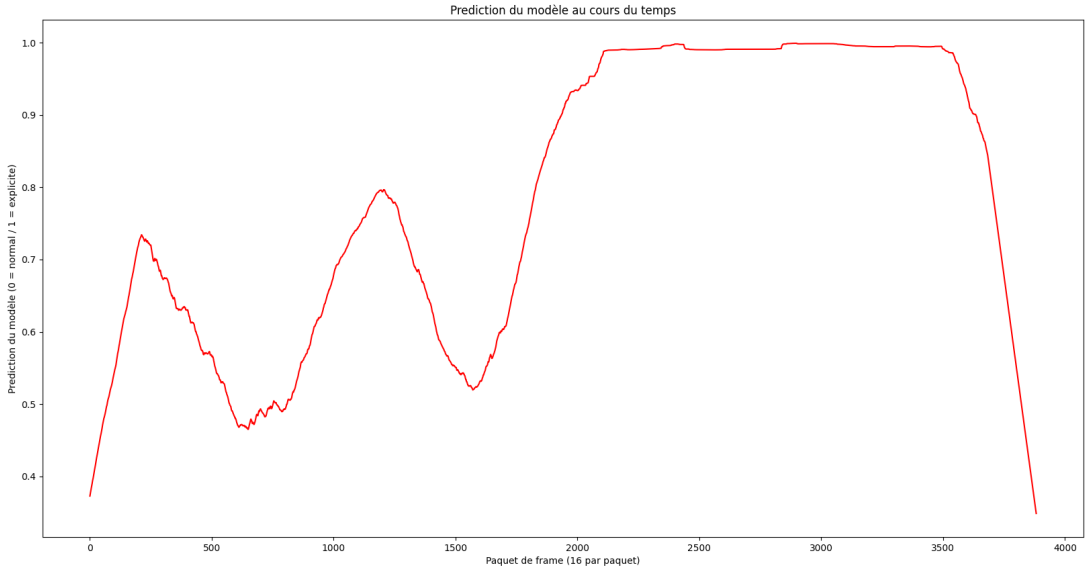

FIGURE 3 – *Illustration d'une analyse des frames d'une vidéo.*

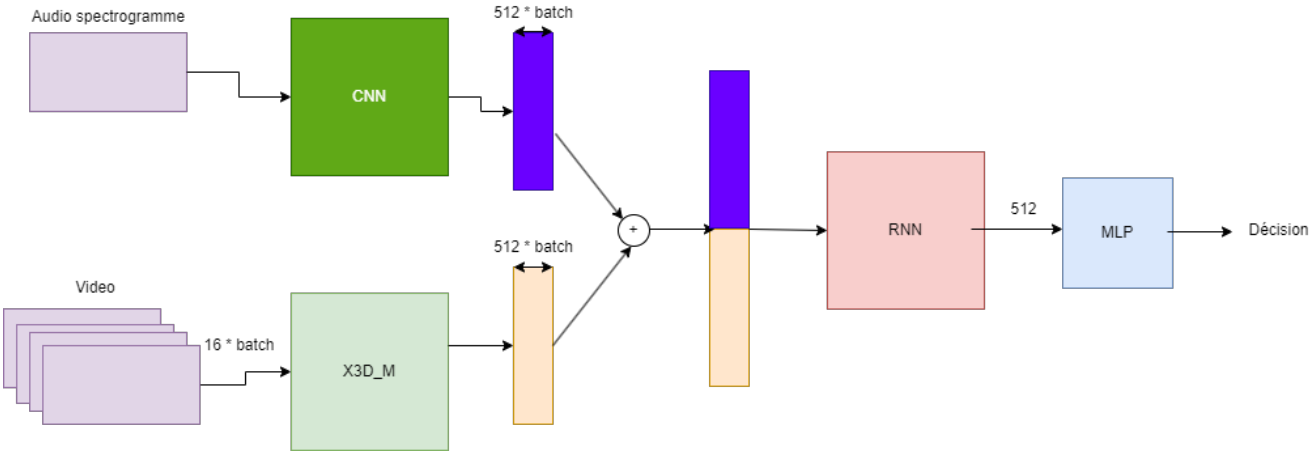

FIGURE 4 – *Architecture du modèle combinant l'audio et la vidéo.*

[3] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, et Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *CoRR*, abs/1711.11248, 2017.

[4] Truc Nguyen et Franz Pernkopf. Lung sound classification using co-tuning and stochastic normalization, 2021.

[5] Konrad Gadzicki, Razieh Khamsehashari, et Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. Dans *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–6, 2020.