Reconnaissance et fouille dans une grande base de vidéos endoscopiques quasi-non annotée

Le candidat:	Tong Yu
Directeur de thèse:	Prof. Nicolas Padoy, PhD
Co-directeur de thèse:	Prof. Didier Mutter, MD
Laboratoire:	CAMMA, AVR, ICube
École doctorale:	ED269 MSII, Université de Strasbourg, France

1 Introduction

En l'espace de quelques années, les techniques de vision par ordinateur ont progressé de manière considérable grâce aux récents développements dans le domaine de l'apprentissage profond. Ces techniques, désormais capables de rivaliser avec la vision humaine sur un grand nombre de tâches visuelles, permettent d'offrir une assistance en temps réel dans de nombreux domaines d'application où les enjeux sont importants, la marge d'erreur est faible et la cadence de travail est soutenue.

Un tel système d'assistance, construit à partir d'algorithmes de vision performants, serait donc particulièrement adéquat au sein d'un bloc opératoire pour venir en aide au chirurgien et au personnel clinique. En particulier, le flux vidéo intra-abdominal utilisé en chirurgie minimalement invasive est une source d'information extrêmement riche, dont le contenu peut être analysé pour fournir au praticien en temps réel des informations pertinentes sur l'action ou l'événement en cours de réalisation. Cependant, un obstacle majeur au développement de ces algorithmes est la rareté des données annotées; entraîner un réseau de neurones de manière complètement supervisée en requiert une quantité considérable.

La base de données dont nous disposons est vaste: plus de 4800 heures d'enregistrements complets d'interventions, sur 12 types de procédures mini-invasives en chirurgie abdominale. Les annotations sont cependant fortement limitées: sur les 518 cholecystectomies et 318 pontages gastriques disponibles, respectivement 120 et 40 sont annotées avec leur décomposition en phases. En l'absence d'annotations, cette base de donnée est pour l'essentiel inexploitée et inexplorée. Collecter plus d'annotations semble être une solution évidente; à grande échelle cependant celle-ci entraînerait des coûts et des délais prohibitifs, d'où la nécessité de solutions alternatives.

Ce constat est le point de départ des travaux présentés dans cette thèse, où nous cherchons à utiliser et explorer cette immense réserve de données brutes pour obtenir des modèles temps réel performants sur des tâches d'analyse vidéo en chirurgie minimalement invasive.

2 Méthodes

2.1 Annotations automatiques pour l'identification temps réel des phases

Une première approche pour naviguer au sein d'une base de données brutes, en l'absence d'annotateurs experts humains, consiste à remplacer ces derniers par un modèle entraîné sur un jeu de données réduit. Nous nous intéressons ici à l'un des problèmes fondamentaux de l'analyse

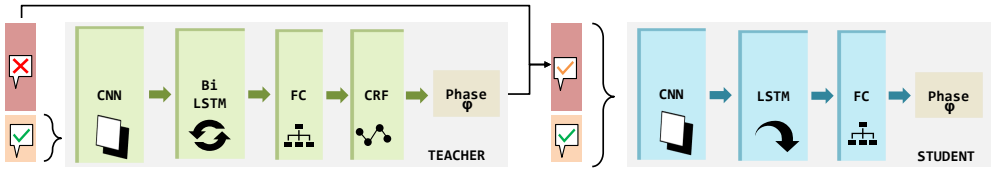

Figure 1: Le modèle CNN-biLSTM-CRF (*maître*) annote des données pour le CNN-LSTM (*élève*), améliorant ainsi ses performances en reconnaissance de phase chirurgicale.

de vidéo chirurgicales, à savoir la reconnaissance des phases. Conformément à notre hypothèse de données annotées rares, nous procèdons de manière semi-supervisée: sur 80 vidéos, nous nous autorisons à utiliser les annotations manuelles sur seulement 1 à 20 vidéos. Les modèles utilisés dans ce problème sont sujets à un compromis: temps réel avec performances limitées (CNN-LSTM), ou a posteriori avec performances maximales (CNN-biLSTM-CRF introduit dans nos travaux).

Nous parvenons à dépasser ce compromis en utilisant les deux types de modèle de manière complémentaire (figure 1): le modèle a posteriori, dans le rôle du *maître*, est entraîné sur les 1 à 20 vidéos manuellement annotées, puis annote automatiquement le reste des 80 enregistrements. L'ensemble sert à entraîner le modèle temps réel, qui tient le rôle d'*élève*. Le F1-score ainsi obtenu est de 78.2 %, contre 70.2 % avec les 20 vidéos manuellement annotées sans annotations automatiques supplémentaires.

2.2 Fouille temps réel de vastes bases de données vidéo génériques

Obtenir des informations sur du contenu vidéo à partir d'un classifieur entraîné à reconnaître un nombre prédéterminé de classes est l'approche canoniquement préconisée en vision par ordinateur, qui pourrait être qualifiée de *directe* et *explicite*. À l'inverse, se renseigner sur un contenu en l'utilisant pour rechercher les éléments pertinents d'une vaste base de donnée est une approche indirecte et implicite, qui propose davantage de flexibilité et de polyvalence.

Les fonctions de hachage profondes constituent une méthode efficace pour la fouille de bases de données; celles-ci calculent des clés binaires de taille réduite à partir de contenus de très grandes dimensions. Ces clés étant optimisées pour préserver les similarités et différences entre contenus, il est alors possible de rapidement effectuer des requêtes par simples comparaisons binaires. Les fonctions de hachage sont souvent entraînées de manière auto-supervisée; leur fonctionnement est dans ce cas complètement indépendant des annotations, qui servent uniquement à leur évaluation quantitative. Le hachage présente donc un double intérêt dans notre situation, en permettant d'utiliser les données non annotées et de les explorer via des requêtes.

La vidéo est cependant fortement en retard par rapport aux autres modalités de recherche telles que les images statiques concernant le hachage. Les quelques travaux traitant de ce problème pour les vidéos ne considèrent de plus que la recherche a posteriori, proposant des méthodes incompatibles avec des recherches temps réel à partir de flux vidéo instantanés.

Le cas temps réel (figure 2) est justement celui que nous choisissons de traiter: considérant une vidéo de durée totale T, nous souhaitons rechercher au sein d'une base de données les vidéos les plus similaires, en observant uniquement les αT premières secondes pour $\alpha \in]0, 1]$ - cette situation correspond au cas d'une diffusion en direct, ou du *streaming*. Pour cela nous

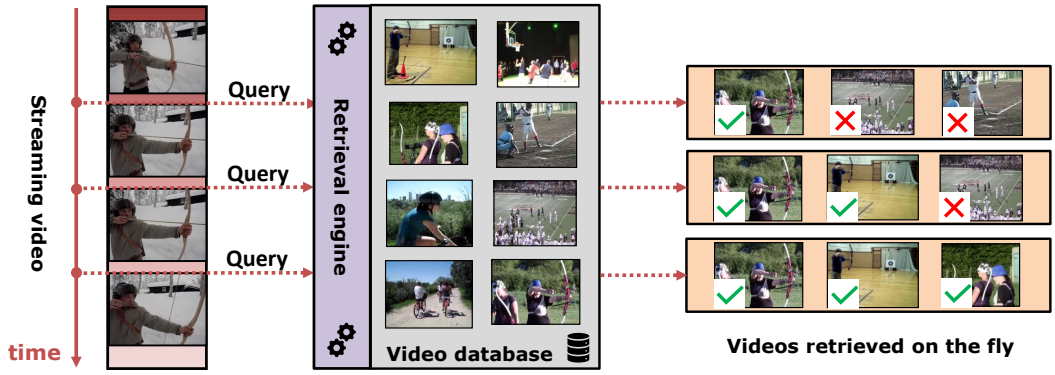

Figure 2: Notre technique de hachage permet de fouiller à l'intérieur de bases de données vidéo de manière incrémentale, au fur et à mesure du déroulement en direct de la vidéo utilisée comme requête. Un archer est filmé; le flux vidéo est soumis au moteur de recherche, qui retrouve deux résultats corrects sur trois

proposons un encodeur vidéo binaire incrémental et prédictif: un modèle LSTM met à jour la clé binaire toutes les deux secondes environ afin que celle-ci tienne compte du contenu visuel en temps réel. Ce modèle est de plus entraîné à anticiper le futur contenu vidéo, afin de compenser le manque d'information visuelle dû à la diffusion en direct.

Nous présentons notre méthode sur deux jeux de données vidéo génériques de très grande taille: FCVID et ActivityNet. L'évaluation s'effectue à partir des étiquettes d'activité fournis par ces jeux de données: dans le classement des meilleurs résultats de recherche, un résultat est considéré correct si sa classe correspond à celle de la requête. Notre méthode, LA-CODE, surpasse les méthodes non prédictives avec notamment 27.4 % de mAP@20 sur FCVID, 14.7 % sur ActivityNet avec moins de 30% de la vidéo observés, pour des clés binaires de 128 bits.

2.3 Fouille temps réel de vastes bases de données vidéo chirurgicales

Pour mettre à l'épreuve nos méthodes de fouilles de données vidéo sur du contenu chirurgical, nous introduisons le jeu de données Laparotome, contenant 2071 enregistrements complets de 12 types de procédures minimalement invasives. Ces enregistrements sont divisés en clips de 30 secondes environ, formant une base de données de plus de 500000 clips.

Nous montrons qu'un même modèle, entraîné sur un vaste ensemble de clips non annotés, permet de rechercher du contenu au sein de trois bases de données: cholecystectomie, pontage gastrique et répertoire d'événements chirurgicaux critiques. L'évaluation s'effectue à partir des phases pour les deux premières; pour la dernière, à partir du type d'événement.

Nous testons deux propriétés utiles pour le fonctionnement temps réel au sein du bloc opératoire: dans un premier temps, la réactivité ou anticipation court terme du hachage, en observant les résultats de recherche lorsque l'observation est limitée aux 10 ou 20 premières secondes. La mAP@10 obtenue avec 128 bits sur les événements critiques est de 28.8 % à 10 secondes contre 36.6% pour un clip complet.

Nous nous intéressons également à la prédictivité, ou anticipation long terme du hachage. Pour cela nous introduisons la méthode dite d'*observation du futur par proxy*. Il s'agit d'une recherche en trois temps: recherche du meilleur proxy dans une base de données parallèle, avance rapide du proxy dans le temps et recherche dans la base de données principale à l'aide du proxy avancé. Cette technique permet de repérer à l'avance les événements susceptibles de

se produire dans les cinq minutes suivantes au milieu d'une chirurgie. La mAP@10 obtenue avec 128 bits sur les événements critiques varie de 19.6 % pour 5 minutes d'avance à 27.1 % pour 1 minute d'avance.

Nous terminons notre étude par une comparaison avec un classifieur conventionnel, en utilisant le modèle de fouille de données en tant que classifieur, avec les résultats de recherche obtenus servant de votes pour la classe à inférer.

3 Conclusion

Les méthodes proposées permettent de tirer parti avec succès des données vidéo endoscopiques non annotées, permettant ainsi de les annoter automatiquement pour l'entraînement de classifieurs temps réel, de les utiliser pour l'entraînement auto-supervisé de modèles pouvant se substituer à des classifieurs, de les explorer par le biais de requêtes vidéo instantanées et enfin d'anticiper les événements futurs d'une chirurgie endoscopique. Ce succès est encourageant pour un passage à l'échelle à partir d'annotations raréfiées, ainsi que le déploiement dans le bloc opératoire de ces méthodes temps réel pour assister le chirurgien.

Publications

[1] T. Yu, D. Mutter, J. Marescaux, and N. Padoy, "Learning from a tiny dataset of manual annotations: A teacher/student approach for surgical phase recognition," *IPCAI*, 2019.

[2] T. Yu and N. Padoy, "Encode the unseen: Predictive video hashing for scalable mid-stream retrieval," *ACCV*, 2020.

[3] C. Nwoye, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy, "Recognition of instrument-tissue interactions in endoscopic videos via action triplets," *MICCAI*, 2020.

[4] S. Ramesh, D. Dall'Alba, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, P. Fiorini, and N. Padoy, "Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures," *IPCAI*, 2021.