Détection d'anomalies dans une séquence d'images fish-eye

Olivier LAURENDIN[1], Sébastien AMBELLOUIS[2], Anthony FLEURY[3], Sanaa CHAFIK[1], Ankur MAHTANI[1]

[1]IRT Railenium
180 rue Joseph Louis Lagrange 59308 Valenciennes Cedex, France

[2]Université Gustave Eiffel, COSYS, LEOST
20 Rue Élisée Reclus, 59650 Villeneuve-d'Ascq, France

[3]IMT Nord Europe, CERI SN
764 Boulevard Lahure, 59500 Douai, France

olivier.laurendin@railenium.eu, sebastien.ambellouis@univ-eiffel.fr,
anthony.fleury@imt-nord-europe.fr,sanaa.chafik@railenium.eu,ankur.mahtani@railenium.eu

Résumé – La détection d'anomalies dans les flux vidéo est un sujet très actif dans la communauté de vision par ordinateur et notamment dans la tâche d'automatisation du traitement de données de vidéo-surveillance. Ce papier propose d'adapter une architecture neuronale fondée sur l'utilisation d'un réseau totalement convolutif utilisant une approche antagoniste générative (réseau GAN) capable d'apprendre la corrélation entre l'apparence des objets présents dans une scène et leurs mouvements apparents et de reconstruire le contenu des images d'une situation qualifiée de normale. L'objectif final est de déterminer la fonction de décision capable d'identifier des situations anormales à partir d'une reconstruction imparfaite d'une situation inconnue. Nous étudions cette problématique dans le contexte de l'automatisation d'un véhicule ferroviaire (projet Train Autonome) pour lequel la fermeture des portes doit se faire en sécurité sans l'intervention d'aucun agent. Cet article propose un ensemble de modifications de l'apprentissage de l'architecture afin de l'adapter aux différents scénarios d'anomalies de notre cas d'application, sensiblement différents de ceux rencontrés dans les datasets classiques de la littérature, et définit les bonnes pratiques pour ce type d'applications. L'architecture est évaluée sur notre dataset nommé « FRailTRI20_DOD » qui met en scène plusieurs exemples d'événements anormaux joués sur la réplique d'une porte de train en laboratoire.

Abstract – Anomaly detection in video streams is a very active subject in the computer vision community and in particular for developing smart video surveillance system. This paper proposes to adapt a neural architecture based on the use of a totally convolutional network within a generative adversarial framework (GAN network) capable of learning the correlation between the appearance of objects present in a scene and their apparent motion and to reconstruct the content of the images of a normal situation. The final objective is to determine the decision function capable of identifying abnormal situations from an imperfect reconstruction of an unknown situation. We study this problem in the context of the automation of a railway vehicle (Autonomous Train project) whose doors closing must be safe without any on-train staff. This article proposes a set of modifications to the learning architecture in order to adapt it to our use-case anomaly scenarios, very different from what is encountered in classic datasets of the litterature, and defines good practices for this type of application. The architecture is evaluated on our dataset named "FRailTRI20_DOD" which features several examples of abnormal events played on a train doors replica in a laboratory.

1 Introduction

Dans ce travail, nous proposons d'identifier les événements anormaux dans une séquence d'images. Ces événements sont généralement très rares tout particulièrement lorsqu'ils sont reliés à la sécurité des usagers d'un système de transport collectif. Face à la difficulté d'observer ces anormalités, il est possible d'adopter des algorithmes de détection dits semi-supervisés dont l'approche consiste à définir les instances dangereuses comme des aberrations par rapport aux instances normales [1]. Les approches neuronales proposent alors des architectures capables d'apprendre à reconstruire le contenu des images d'un scénario normal en faisant l'hypothèse qu'en présence d'un événement anormal jamais observé l'erreur de reconstruction sera plus importante et permettra de les détecter. Dans notre application, ces approches sont intéressantes puisque les événements anormaux étant très rares et imprévisibles, prédire tous les événements possibles à proximité des portes de train est une tâche ardue nécessitant des ressources importantes tant pour leur acquisition que pour leur annotation nécessaire à un entraînement supervisé. Les annotations d'anomalies nécessaire à une approche semi-supervisée se limitent uniquement aux données utilisées pour évaluer les performances du modèle, les données d'entraînement n'étant constituées que de données dénuées d'anormalité, très abondantes et ne nécessitant aucune annotation.

Plusieurs bases de détection d'anomalies mettent en scène des piétons (UCSD [2], UMN [3]). Des bases sont également spécifiques au contexte ferroviaire tels que PAMELA-UANDES [4], BOSS [5] et FRailTRI20_DOD [6], que nous avons acquise dans une interface train-gare reconstituée. Les scénarios

normaux consistent en l'ouverture et la fermeture d'une reproduction de portes de train et le franchissement de l'embrasure. Les événements anormaux sont caractérisés par l'interruption de la fermeture des portes dû à un défaut mécanique, de piétons/bagages coincés entre les portes ou de chutes de passagers. Tout comme BOSS et PAMELA-UANDES, ces images sont prises à partir d'une caméra fish-eye placée sur le plafond du train (cf. fig. 1).

Notre étude porte sur les architectures dont l'erreur de reconstruction est la combinaison d'une erreur d'apparence et d'une erreur de mouvement [7][8]. Ces architectures exploitent des auto-encodeurs ou des GANs et utilisent généralement une image d'entrée et le flux optique correspondant (cf fig. 2). Parmi les solutions proposées dans la littérature, nous avons retenu le travail de Nguyen et al. [9], capable de saisir les informations de plusieurs instances simultanément.

Là où la plupart des anomalies présentes dans d'autres bases sont de nature ponctuelle, soit par l'apparition d'une instance inconnue, soit par un changement soudain du profil de mouvement d'une instance[2],[3]. Les anomalies de notre base sont plus complexes (fig. 1). Le cas des piétons coincés entre les portes par exemple est le résultat d'une interaction anormale porte/piéton. En entraînant sur notre base un réseau conçu initialement pour d'autres bases et en analysant les résultats, nous cherchons à voir dans quelle mesure nos événements dangereux peuvent être résumés à un ensemble d'anomalies ponctuelles en apparence ou en mouvement.

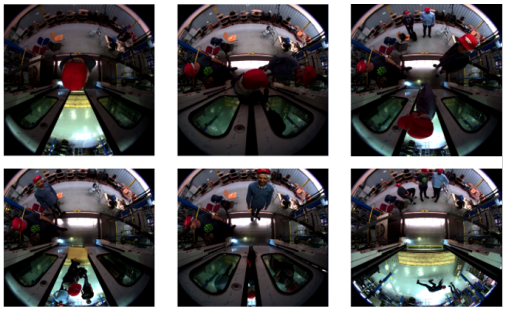

FIGURE 1 – Images anormales de la base « FRail-TRI20_DOD »

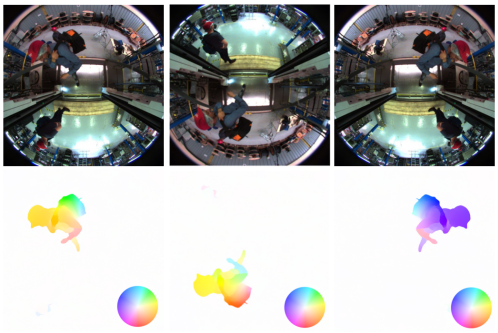

FIGURE 2 – Image et flux optique associé. À partir des données en colonne 1, on effectue une symétrie verticale et horizontale pour obtenir les colonnes 2 et 3. La teinte du flux optique indique sa direction et la saturation sa norme.

2 Méthode proposée

2.1 Réseau d'origine

Pour prédire un score d'anomalie pour une image donnée en entrée au réseau de Nguyen et al [9], celle-ci est d'abord fournie à un auto-encodeur U-Net (générateur). Il est composé d'un encodeur suivi de deux décodeurs. L'encodeur compresse l'image originale qui est décompressée par les deux décodeurs. Le premier décodeur est entraîné à reconstruire l'image originale tandis que le second en déduit le flux optique. L'erreur pour l'estimation de l'anormalité de l'image d'entrée est une combinaison d'une erreur de reconstruction en apparence, entre l'image d'entrée et l'image reconstruite, et d'une erreur de mouvement, entre le flux optique prédit et le flux optique de vérité terrain fourni par un réseau tiers.

Pour aider le deuxième décodeur à déduire le flux optique de l'image d'entrée, un discriminateur supplémentaire est entraîné à faire la distinction entre le flux optique prédit par le générateur et le flux optique vérité terrain en utilisant l'image d'entrée comme référence. Le discriminateur fournit des indications, sous la forme d'une fonction de perte antagoniste, au générateur sur les caractéristiques qui distinguent le flux optique prédit de sa vérité terrain. La fonction de perte utilisée pour entraîner le générateur est donc la somme pondérée de trois fonctions de perte : une perte en apparence, une perte en flot optique et une perte antagoniste.

$$PSNR(x,y) = 10log\left(\frac{d^2}{(x-y)^2}\right) \quad (1)$$

$$SSIM(x,y) = \frac{2\mu_x\mu_y}{\mu_x^2\mu_y^2}\frac{2\sigma_x\sigma_y}{\sigma_x^2\sigma_y^2}\frac{cov_{xy}}{\sigma_x\sigma_y} \quad (2)$$

d : dynamique de l'entrée, μ, σ et cov : moyenne, variance et covariance.

2.2 Modifications proposées

2.2.1 Fonction de perte pour le flot optique

La norme du flot optique moyen d'une même instance est plus élevée lorsqu'elle se trouve au centre de l'image. L'erreur de reconstruction est donc plus impactée lorsque l'instance est au centre de l'image que lorsqu'elle s'en éloigne. Nous expérimentons donc un ensemble de fonctions de perte pour le flux optique pour aider le réseau à se concentrer sur les parties pertinentes de l'image. La première série d'approches consiste en des heuristiques pour ignorer les parties statiques de l'image en calculant la valeur moyenne de l'erreur en norme L1 sur le flux optique restreinte à certaines zones du plan image (éq. 4). En définissant un masque circulaire (Radius Mask) nous réduisons la zone du calcul d'erreur aux pixels dont la distance au centre de l'image est inférieure à un rayon donné. Nous utilisons aussi

des masques qui réduisent la zone d'intérêt aux pixels dont la norme du flux optique de la vérité terrain (Simple Norm Mask) ou celle la somme de la vérité terrain et du flux optique prédit (Symmetrical Norm Mask) est supérieure à un seuil donné. La deuxième série d'approches consiste à remplacer la norme L1 par la fonction de perte de Ruzicka (éq. 5)[10], i.e. la perte quantitative de Jaccard. Pour une paire de masques à valeurs flottantes donnée, cette perte est minimale lorsque ces masques correspondent parfaitement et ignore les pixels de valeurs simultanément proches de zéro dans les deux masques. Cependant, comme cette perte n'est définie que pour des masques à valeurs positives, elle ne peut être appliquée que sur la norme du flot optique (éq. 6).

$$L1_{msk}(x,y) = \sum_{msk} |x-y| \otimes msk \quad (3)$$

$$L1_{\mathbf{v}} = L1_{msk}(\mathbf{v}_{vt}, \mathbf{v}_{pred}) \quad (4)$$

$$Ruzicka(x,y) = 1 - \frac{min(x,y)}{max(x,y)} \quad (5)$$

$$Ruzicka_{\mathbf{v}} = Ruzicka(|v_{vt}|, |v_{pred}|) + L1_{|v_{vt}|>\epsilon}(\theta^v{}_{vt}, \theta^v{}_{pred}) \quad (6)$$

avec v_{vt} et v_{pred} vecteurs de mouvement de la vérité terrain et predit, θ_{vt} et θ_{pred} la direction d'un vecteur de mouvement, $\otimes$ produit de Hadamard, ϵ seuil fixé et msk un masque de pixels.

2.2.2 Apparence dépendante de la position dans l'image

Une autre conséquence de l'utilisation d'une caméra fish-eye positionnée au plafond est que l'apparence d'une instance donnée n'est plus invariante par translation. Cette propriété va à l'encontre des hypothèses de base pour l'utilisation des réseaux de neurones convolutifs, invariants par translation. Nous voulons que le réseau soit capable de distinguer un piéton en position verticale dans la moitié supérieure et dans la moitié inférieure. Un piéton debout serait tête en haut dans le premier cas, tête en bas dans le second (cf. fig. 2). La question de la rupture de la propriété d'invariance par translation des réseaux de neurones convolutifs a été abordée dans [8] avec l'utilisation d'une couche appelée *CoordConv*. Elle consiste à ajouter les coordonnées des pixels en entrées aux filtres convolutifs du réseau. Nous expérimentons deux couches *CoordConv* comme première couche de notre réseau en coordonnées euclidiennes et polaires.

2.2.3 Augmentation de données

Grâce à la symétrie des images, nous avons accru artificiellement la diversité de nos données pour pallier au manque de données de la tâche étudiée en retournant aléatoirement images et flux optiques horizontalement ou verticalement (fig. 2).

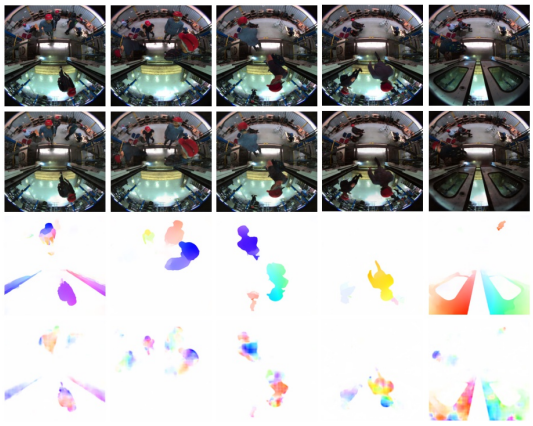

FIGURE 3 – Résultats avec le réseau avec masque de Ruzicka symétrique. Lignes de haut en bas : image d'entrée, image reconstruite, flux vérité terrain, flux prédit.

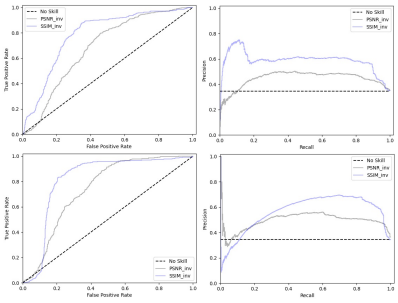

FIGURE 4 – Courbe ROC et courbe précision-rappel pour le réseau avec masque Ruzicka symétrique avec erreur en apparence en haut et en mouvement en bas.

3 Résultats expérimentaux

Les réseaux sont entraînés sur 200 epochs avec un batch de taille 8, en utilisant l'optimiseur Adam avec taux d'apprentissage initiaux de 2.10^{-4} (générateur) et 2.10^{-5} (discriminateur). Les résultats sont présentés dans le tableau 1 pour l'AUC-ROC et pour la précision et le rappel moyens (aPR), pour les erreurs en apparence et en mouvement, selon le PSNR (éq. 1) et la SIMilarité Structurelle (SSIM, éq. 2). Les données prédites et la vérité terrain sont standardisées pour de meilleurs résultats. Les résultats d'un lancé aléatoire en AUC-ROC et en aPR et de l'implémentation originale sont fournis. Certains résultats

	AUCROC				aPR			
	Apparence		Mouvement		Apparence		Mouvement	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Aléatoire	0.5				0.33			
Original	0.659	0.707	0.741	0.667	0.417	0.454	0.544	0.415
Radius Mask	0.566	0.807	**0.805**	0.796	0.353	0.650	**0.610**	0.536
Norm Mask (Simple)	**0.711**	**0.828**	0.566	0.764	**0.504**	**0.679**	0.404	0.505
Norm Mask (Sym)	0.690	0.770	0.720	0.799	0.462	0.549	0.541	0.534
Ruzicka Mask (Simple)	0.635	0.791	0.767	0.812	0.414	0.623	0.532	**0.562**
Ruzicka Mask (Sym)	0.637	0.785	0.768	**0.816**	0.426	0.587	0.529	0.560
Data augmentation	0.585	0.748	0.725	0.658	0.388	0.531	0.499	0.398
Coordconv (euclidian)	0.480	0.817	0.768	0.806	0.312	0.652	0.556	0.538
Coordconv (polar)	0.480	0.810	0.800	0.805	0.324	0.629	0.604	0.537

TABLE 1 – Résultats expérimentaux

pour le masque symétrique de Ruzicka sont donnés fig. 3 et les courbes ROC et précision-rappel sont présentées fig. 4.

L'analyse du tableau montre que la détection des anomalies est en moyenne meilleure selon SSIM, notamment en apparence. Le PSNR étant une mesure à l'échelle du pixel (versus à l'échelle d'un patch pour SSIM), cela indique la présence d'un bruit local qui nuit à la reconstruction en apparence. Conformément à [6], les événements dangereux sont plus faciles à distinguer à partir du mouvement que de l'apparence. Il faut noter que le réseau est souvent meilleur pour prédire la norme du flot optique que sa direction. Comme montré fig. 3, le réseau prédit les parties mobiles de l'image (piétons, portes) mais il est plus difficile d'identifier la direction des mouvements. Toutefois, il est en mesure de prédire la direction du mouvement des portes avec une seule image, ce qui constitue un résultat intéressant.

Le masque Radius apporte une meilleure détection d'anomalie de mouvement mais induit une perte partielle de performance en anomalie d'apparence (PSNR). Un masque de norme simple augmente les performances sauf en détection d'anomalies de mouvement (PSNR). La variante symétrique lisse les résultats. Cela semble indiquer qu'une partie du bruit induit par l'utilisation du masque simple est atténuée par la prise en compte du flux optique prédit. Ruzicka permet un gain de performance indépendamment de la variante utilisée et montre sa capacité à se concentrer sur les parties mobiles. L'augmentation des données n'améliore pas les performances. L'utilisation de *CoordConv* en entrée a un impact majeur sur la détection des anomalies de mouvement et des anomalies d'apparence en termes de SSIM.

4 Conclusion

Après analyse des particularités de nos données, une série de modifications à apporter à un algorithme de détection d'anomalies de la littérature est proposée. L'utilisation de masques sur le flot optique permet au réseau de se concentrer sur les parties mobiles de l'image, parfois au détriment de leur reconstruction. Cette perte de performance en terme de reconstruction peut être atténuée en incorporant le flot optique prédit dans la fonction de perte avec un masque symétrique ou en utilisant la fonction de perte de Ruzicka. La rupture de l'invariance par translation des filtres convolutifs aide le réseau à identifier les événements anormaux en tenant compte de la position dans l'image.

Remerciements

Ce travail de recherche est financé par le programme français "Investissements d'Avenir" et il s'inclut dans le projet collaboratif français TASV (Train Autonome Service Voyageurs), avec Railenium, SNCF, Alstom Crespin, Thales, Bosch, et SpirOps.

Références

[1] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection : A review," *ACM Comput. Surv.*, vol. 54, no. 2, mar 2021.

[2] V. Mahadevan, W.-X. LI, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1975–1981.

[3] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[4] S. A. Velastin and D. A. Gómez-Lira, "People Detection and Pose Classification Inside a Moving Train Using Computer Vision," in *Advances in Visual Informatics*, ser. Lecture Notes in Computer Science, B. Z. et al., Ed. Cham : Springer International Publishing, 2017, pp. 319–330.

[5] C. Lamy-Bergot, S. Ambellouis, L. Khoudour, D. Sanz, N. Malouch, A. Hocquard, J.-L. Bruyelle, L. Petit, A. Cappa, A. Barro, E. Villalta, G. Jeney, and K. Egedy, "Transport system architecture for on board wireless secured A/V surveillance and sensing," in *2009 9th International Conference on Intelligent Transport Systems Telecommunications, (ITST)*, Oct. 2009, pp. 564–568.

[6] O. Laurendin, S. Ambellouis, A. Fleury, A. Mahtani, S. Chafik, and C. Strauss, "Hazardous events detection in automatic train doors vicinity using deep neural networks," in *2021 17th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2021, pp. 1–7.

[7] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7834–7843, 2019.

[8] W. Liu, W. Luo, D. Lian, and S. Gao, "Future Frame Prediction for Anomaly Detection - A New Baseline," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 6536–6545, iSSN : 2575-7075.

[9] T. N. Nguyen and J. Meunier, "Anomaly Detection in Video Sequence With Appearance-Motion Correspondence," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 1273–1283, iSSN : 2380-7504.

[10] M. J. Warrens, "Inequalities Between Similarities for Numerical Data," *Journal of Classification*, vol. 33, no. 1, pp. 141–148, Apr. 2016. [Online]. Available : http ://link.springer.com/10.1007/s00357-016-9200-z