Université de Paris
École doctorale EDITE (130)
Laboratoire d'Informatique Paris Descartes (EA 2517)

Prise en compte de l'information spatiale et temporelle pour l'analyse de séquences d'images

Présentée et soutenue publiquement le 26-11-2021 par :

Mohamed Tayeb Chelali

Thèse de doctorat en :

Informatique

Spécialité :

Imagerie

Membres du jury

Jenny Benois-Pineau	Professeure, Université de Bordeaux	Rapporteure
Sébastien Lefèvre	Professeur, Université de Bretagne Sud	Rapporteur
Marin Ferecatu	Maître de conférences (HDR), CNAM	Examinateur
Camille Kurtz	Maître de conférences, Université de Paris	Examinateur
Serge Miguet	Professeur, Université Lumière Lyon 2	Examinateur
Nicolas Passat	Professeur, Université de Reims Champagne-Ardenne	Examinateur
Anne Puissant	Professeure, Université de Strasbourg	Examinatrice
Nicole Vincent	Professeure, Université de Paris	Directrice de thèse

Remerciements

Je tiens à remercier en premier lieu pleinement mon encadrant Camille Kurtz et ma directrice de thèse Nicole Vincent qui ont su me guider et diriger le long de la thèse. Ils se sont donnés à fond pour me fournir un climat de travail confortable et satisfaisant aussi bien au niveau matériel et intellectuel. *Camille* est comme mon « grand frère » et *Nicole* est comme ma « mère » dans le monde de l'enseignement et de la recherche. Durant toute la période de cette thèse, ils ont été là pour répondre à mes questions et ont su m'encourager pour me pousser à m'améliorer dans la recherche et l'enseignement.

Un grand merci à *Anne Puissant* pour ses contributions dans l'encadrement de cette thèse. Son expertise de géographe en analyse de l'environnement a été un plus pour la compréhension et la mise en place des besoins opérationnels des images satellitaires.

J'adresse mes remerciements à *Jenny Benois-Pineau* et *Sébastien Lefèvre* pour leur relecture de ce manuscrit de thèse et l'intérêt qu'ils ont porté à mes travaux de recherches. Je remercie également *Serge Miguet*, *Nicolas Passat*, *Marin Farecatu* d'avoir accepté de participer à mon jury de thèse.

Mes remerciements vont à toutes les équipes du projet TIMES. Les échanges que nous avons eus entre nous ont été enrichissants et m'ont permis d'évoluer et d'approfondir mon savoir dans le domaine de la géographie.

Je ne saurai oublier les membres de l'équipe SIP avec qui j'ai passé trois ans de travail. Je les remercie tous, en particulier *Florence Cloppet*, *Nicolas Loménie* et *Laurent Wendling* qui ont été des références pour moi. Un grand merci aux doctorants, post-doctorants avec qui j'ai passé des moments agréables au laboratoire, en particulier *Héloïse*, *Olivier*, *Robin*, *Thibault*, *François*, *Guillaume* et *Christian*.

Je tiens à remercier mes amis qui m'ont encouragé à aller de l'avant, en particulier *Warith*, *Ismet*, *Walid* et *Anis*.

Enfin, un plus grand « merci » à mes parents et mes frères, *Rabah* et *Yahia*, qui m'ont toujours soutenu et supporté. Leurs aides et encouragements étaient très précieux dans mes choix, parfois difficiles, tout au long de mon parcours universitaire.

Table des matières

Table des figures

Liste des tableaux

Résumé

L'évolution de la technologie numérique a permis la multiplicité des capteurs d'images avec lesquels des masses de données visuelles sont quotidiennement produites. Dans certains contextes, ces données peuvent prendre la forme de séquences temporelles d'images $2D$ conduisant à des données $3D$ que nous noterons $2D + t$. Ce type de données est fréquent dans plusieurs domaines tels que la télésurveillance ou la télédétection. De par leur dimension, l'analyse et l'interprétation de toute cette masse de données constitue un des défis importants en vision par ordinateur. Cette thèse s'inscrit dans le contexte de l'exploitation de ces données afin de pouvoir les classifier, en exploitant au maximum la richesse des informations spatiales et temporelles portées par ces données. Les travaux de recherche présentés dans ce manuscrit comprennent deux méthodes qui procèdent différemment mais dont le point commun repose sur un changement de représentation des données initiales. La première méthode se base sur l'extraction de caractéristiques expertes (*hand-crafted*) tandis que la deuxième concerne l'utilisation des méthodes d'apprentissage automatique, en particulier les réseaux de neurones convolutifs profonds. À travers ces deux méthodes, nous nous proposons d'étudier la stabilité temporelle des séquences temporelles d'images avec les caractéristiques expertes et étudier leurs variabilités spatiale et temporelle avec les réseaux de neurones convolutifs profonds. Les deux méthodes sont ensuite évaluées sur deux applications différentes. Une de ses applications concerne les séries temporelles d'images satellitaires et l'autre concerne les vidéos de caméra de surveillance. Les résultats expérimentaux illustrent l'intérêt des méthodes proposées.

Abstract

The evolution of digital technology has allowed the multiplicity of image sensors, leading every day to the production of masses of visual data. In some contexts, these data can take the form of $2D$ images time series leading to $3D$ data that we note $2D + t$. This type of data is frequent in several domains such as remote surveillance or remote sensing. Because of their dimensions, the analysis and interpretation of this mass of data is a major challenge in computer vision. This thesis is in the context of the exploitation of these data in order to classify them, by exploiting the maximum the wealth of spatial and temporal information carried by these data. The research works presented in this manuscript includes two methods that proceed differently but whose common point is based on a change of the representation of the initial data. The first method is based on the extraction of hand-crafted features while the second one is based on the use of machine learning methods, in particular deep convolutional neural networks. Through these two methods, we propose to study the temporal stability of image times series with hand-crafted features and to study their spatial and temporal variability with deep convolutional neural networks. The two methods are then evaluated on two different applications. One is related to satellite image time series and the other is related to surveillance camera videos. The experimental results illustrate the interest of the proposed methods.

Chapitre 1

Introduction

> Science sans conscience n'est que ruine de l'âme.
>
> – François Rabelais, 1532

De nos jours, la quantité des données visuelles a explosé grâce à la multiplicité des capteurs d'images. Ces données peuvent être sous la forme d'image fixe ou d'une séquence temporelle d'images. Cela a lancé un défi au sein de la communauté de vision par ordinateur dans le contexte de l'analyse et l'exploitation de cette masse de données et ce de façon automatique.

Dans ce chapitre d'introduction, nous présentons dans la section 1.1 le contexte de l'étude menée durant cette thèse de doctorat qui porte sur l'analyse des séquences temporelles d'images. La section 1.2 présente les méthodes classiques qui traitent les séquences d'images. La section 1.3 évoque les problématiques et les difficultés des méthodes existantes. Les sections 1.4 et 1.5 présentent nos motivations et nos contributions pour apporter des solutions aux problèmes évoqués.

1.1 Contexte des travaux

La multiplicité des capteurs d'images conduit chaque jour à la production de masses de données visuelles. Les données peuvent prendre la forme d'images fixes ($2D$) ou de séquences temporelles d'images ($2D + t$) où une même scène est captée à des moments différents. Les séquences d'images peuvent être utilisées à des fins industrielles, pour faire de la recherche scientifique ou même pour les loisirs. Nous pouvons noter deux cas d'applications : la réalisation des films ou l'analyse de mouvement dans les vidéos / les courtes vidéos diverses que nous trouvons dans les réseaux sociaux.

Ce type de données s'est même étendu dans plusieurs autres domaines. En télédétection, les satellites équipés de capteurs optiques ou radars survolent la Terre et prennent régulièrement des images de certaines régions. Les données peuvent être utilisées pour des études environnementales ou la cartographie de la couverture terrestre. Par exemple, la constellation de satellites d'observation de la Terre Sentinel-2 fournit des séquences d'images avec des résolutions spatiales, spectrales et temporelles élevées sur toute la surface du globe [33]. En médecine, les appareils d'imagerie radiologique peuvent être utilisés pour suivre chaque mois l'évolution d'une pathologie chez un patient dans le cadre d'études longitudinales [82]. En biologie, une caméra fixée sur un microscope peut être utilisée pour analyser le développement d'une cellule pendant quelques minutes [127]. En télé-surveillance, les caméras de sécurité permettent le suivi des gens dans les endroits publics [120].

Les communautés spécialisées en vision par ordinateur s'intéressent ainsi au développement d'algorithmes permettant d'analyser, de traiter et de comprendre ces masses de données visuelles de façon automatique et performante. Ces algorithmes font appel à des techniques et stratégies issues du traitement d'images, de l'intelligence artificielle et de l'apprentissage machine. Parmi les principales applications, nous trouvons en général l'étiquetage d'images, la cartographie d'occupation du sol en télédétection, l'extraction de texte dans une image de document et aussi la segmentation sémantique de scènes en analyse d'images de vidéos ou de scanner des cerveaux. La figure 1.1 illustre un exemple de chacune de ces applications. Dans notre cas, nous ne nous limitons pas aux traitements d'une image fixe mais plutôt à celui d'une séquence d'images. Avec de telles données, le traitement devient plus complexe car la dimension temporelle est rajoutée aux deux dimensions spatiales traditionnellement considérées en traitement d'images.

Une des étapes importantes de la vision par ordinateur est l'extraction de caractéristiques à partir des données. Ces caractéristiques sont utilisées dans différents étages de la chaîne d'analyse d'images. Dans notre cas, ces dernières doivent être extraites à partir de séquences d'images. Ces informations peuvent être les valeurs brutes du pixel, des indices de contours ou de textures. Ce traitement est une étape primordiale permettant d'atteindre des résultats très prometteurs dans diverses applications. Dans ce contexte, l'équipe Systèmes Intelligents de Perception (SIP) du Laboratoire d'Informatique Paris Descartes (LIPADE), spécialisée dans diverses problématiques liées à l'analyse et la compréhension d'images, travaille sur l'extraction de caractéristiques dans diverses applications. Parmi ces dernières, nous citons la segmentation d'images, le suivi d'objets en mouvement ou encore la descrip-

Figure 1.1 – Illustration des résultats de quelques applications de la vision par ordinateur. Ces applications sont : la classification d'images, la cartographie de l'occupation des sols, la détection de texte dans des images de documents, la segmentation d'image d'IRM ou la segmentation de scène.

tion des relations spatiales entre objets d'intérêt dans les images. Ces travaux sont menés dans différents domaines tels que l'imagerie satellitaire, l'imagerie biologique et médicale ou l'analyse de documents (manuscrits et imprimés). Cette thèse s'inscrit par ailleurs dans le contexte du projet TIMES [1], financé par l'Agence Nationale de Recherche (ANR) pour une période de 48 mois (Novembre 2017 – Octobre 2021). Elle a donc pour objet la production de nouvelles méthodes sur l'extraction de caractéristiques permettant de comprendre la dynamique des séquences d'images qui peuvent être satellitaires ou issues de vidéos tout en tenant compte de l'information spatiale et temporelle simultanément. Ces caractéristiques seront ensuite utilisées dans différentes tâches d'analyse de séquences temporelles d'images.

1.2 Analyse de séquences d'images

Une séquence d'images, par exemple une vidéo regroupe un ensemble d'images acquises par un ou plusieurs capteurs à des instants différents et ordonnées chronologiquement. Ce

1. Exploitation de masses de données hétérogènes à haute fréquence temporelle pour l'analyse des changements environnementaux (https://anr.fr/Projet-ANR-17-CE23-0015)

type de données contient des informations spatiales et temporelles qui représentent deux aspects relatifs à la scène observée. L'aspect spatial permet d'avoir une information sur la répartition des objets dans l'image tandis que l'aspect temporel permet d'étudier l'évolution temporelle de ces objets (*e.g.*, mouvement). On notera STI comme *Séquence Temporelle d'Images* ces données. Elles sont représentées dans un espace à trois dimensions. Les STI peuvent être encodées sous la forme d'un cube de données à deux dimensions spatiales et une dimension temporelle. L'acquisition d'une STI peut se faire avec un ou plusieurs capteurs afin d'obtenir une séquence d'images plus importante avec une fréquence temporelle élevée.

L'Humain dispose d'une perception permettant d'analyser ces données par photo-interprétation afin de les décrire de façon sémantique, en langage naturel, selon leurs contenus visuels. Face à cette masse de données, la réalisation d'une telle tâche s'avère fastidieuse, voire même impossible dans certains contextes particuliers et l'automatisation du système devient alors indispensable.

L'analyse et l'exploitation de ces données, en prenant en compte les informations spatiales et temporelles, permet la compréhension de certains phénomènes particuliers comme l'évolution d'un objet d'intérêt qui n'est pas observable et analysable à partir d'une seule image de la séquence. Parmi ces phénomènes, on peut citer le développement saisonnier de la végétation analysable à partir de séries temporelles d'images satellitaires [107, 129, 145] ou la réduction d'une tumeur en imagerie médicale [70]. En fonction des applications envisagées, de nombreuses tâches d'analyse d'images et de vision par ordinateur sont effectuées sur ces données telles que la détection, le suivi d'objets d'intérêt, la segmentation, la reconnaissance d'actions, etc. Chacune de ces méthodes consiste à extraire des informations caractérisant la tâche recherchée à partir des STI de façon automatique. Néanmoins, la description du contenu des données par leurs contenus sémantiques relativement aux données visuelles reste difficile. Par exemple, pour une STI qui représente une scène ou un objet d'intérêt particulier, il est nécessaire de prédire une étiquette de classe potentiellement liée à son évolution dans le temps.

De nos jours, l'exploitation des STI peut se faire de façon automatique grâce à l'évolution de la technologie et la disponibilité des moyens de calcul. Mais, les données brutes des STI impactent la qualité des résultats attendus. Par exemple, utiliser uniquement l'évolution des valeurs colorimétriques des pixels à travers le temps reste très limité pour catégoriser un ensemble d'images. Pour cela, l'utilisation des méthodes de traitement d'images devient indispensable afin de récupérer des informations qui caractérisent au mieux les données brutes. Par la suite, ces informations seront donc utilisées pour étiqueter les données, par exemple pour une tâche de classification.

1.3 Problématique

En vision par ordinateur, toute démarche scientifique commence par fixer une problématique en fonction de l'acquisition ou de la récolte des données nécessaires à la base du

image 18 *image* 34 *image* 50 *image* 68

(a) Une STI avec une continuité visuelle.

10/01/2017 10/04/2017 30/05/2017 17/07/2017 21/08/2017 07/10/2017 14/11/2017

(b) Une STI avec des images ponctuelles.

Figure 1.2 – Illustration de deux types de séquences temporelles d'images (STI).

problème. Néanmoins, les données numériques retranscrivant une scène du monde réel contiennent des artéfacts de bruits ou des impuretés qui sont dues au capteur optique et à la luminosité de la scène. Pour cela, l'analyse et la compréhension des données est une étape primordiale à ne pas négliger. De plus, l'utilisation des méthodes automatiques de traitement d'images permettent la mise en évidence des informations caractérisant le contenu des données. Les STI peuvent contenir deux types de contenus qui dépendent de la scène observée. La différence entre ces deux types de données repose sur la continuité visuelle. Le premier type représente des vidéos contenant des objets en mouvement, comme il est présenté dans la figure 1.2.a. Ce type de données présente un exemple dans lequel la continuité visuelle est assurée car il y a au moins 25 images qui passent pendant une seconde. Si il y a moins de 25 images par seconde, la vidéo peut être vue au ralenti avec des déplacements brusques des objets observés. Le deuxième cas correspond à un contenu non-déformable où les pixels de la séquence représentent la même zone spatiale au cours du temps, par exemple les Séries Temporelle d'Images satellitaire ($STIS$) où une même zone est observée dans le temps, la figure 1.2.b illustre ce type de contenu. Dans ce cas, le capteur optique est focalisé sur une seule zone et capte plusieurs images de la scène à des moments ponctuels. Les méthodes d'extraction de caractéristiques traitent ces deux types de données différemment. Les STI avec des objets en mouvement sont traitées de façon à avoir une information sur le mouvement subi spatialement dans le temps. Par contre, c'est davantage l'évolution temporelle qui est étudiée pour les STI avec un contenu non-déformable.

Au cours de notre présente étude, nous nous focalisons sur la classification d'un objet d'intérêt ou d'une scène observée à partir de STI en utilisant des méthodes de traitement d'images. Naturellement les cas où les deux aspects, spatial et temporel, sont significatifs recevront davantage notre attention. La figure 1.2 illustre respectivement en (a) et (b) une personne qui marche et une prairie captée à des moments différents dans l'année (les images sont ordonnées chronologiquement de gauche à droite). La classification de la vidéo peut

se faire par l'analyse de la déformation spatiale de la personne dans le temps. Par contre la prairie ne se déforme pas mais la texture peut être utilisée. Cette dernière reste cependant limitée quand deux types de végétations se ressemblent comme le blé et le seigle. La texture peut aussi être utilisée sur la vidéo, mais étant donné que la personne change seulement de position et non pas d'habits (même information de texture), la texture va alors rester presque similaire entre les différentes images de la séquence. Par contre, elle est davantage utile dans le suivi de personnes qui se croisent. Dans ce contexte, nous nous intéressons aux problèmes de l'extraction de caractéristiques, pas seulement spatiales ou temporelles, mais qui sont nativement spatio-temporelles à partir de ces deux types de contenus.

1.4 Motivations

Les travaux menés dans cette thèse se focalisent sur les méthodes d'extraction de caractéristiques à partir de STI. Deux grandes familles de méthodes d'extraction de caractéristiques existent. La première famille requiert l'intervention d'un expert humain afin de choisir et/ou de définir des caractéristiques adaptées au besoin. Les caractéristiques extraites ici sont fabriquées à la main ou dites en anglais caractéristiques « *hand-crafted* ». La deuxième famille concerne les méthodes basées sur un apprentissage afin d'apprendre des caractéristiques optimisées par rapport au problème ciblé. Ces méthodes dites « *end-to-end* » (bout en bout) ont souvent recours aux réseaux de neurones profonds. Elles sont aussi destinées à être autonomes, c'est-à-dire qu'elles apprennent à partir des données d'entrée.

La nature des caractéristiques dépend du niveau de traitement des méthodes. Il y a celles qui ne considèrent que le domaine temporel et n'amènent à produire que des caractéristiques temporelles. L'aspect spatial des données est totalement ignoré. D'autres méthodes, pour chaque image, produisent des caractéristiques spatiales. Puis une agrégation des informations extraites à travers le temps est appliquée avant la décision finale. À ce niveau, les caractéristiques ne sont que spatiales mais évoluent dans le temps. Néanmoins, dans diverses applications, l'étude conjointe du domaine spatial et du domaine temporel peut permettre une analyse plus fine de la scène observée et une meilleure compréhension des phénomènes spatio-temporels qui peuvent caractériser les objets d'intérêt étudiés et leurs évolutions. Dans ce contexte, certaines approches combinent des caractéristiques spatiales et temporelles pour analyser une STI. Cependant, les deux domaines sont souvent traités individuellement et une fusion est alors opérée au niveau décisionnel. À ce niveau, les caractéristiques ne sont pas purement ou nativement spatio-temporelles. Enfin, les méthodes qui traitent tout le cube de données produisent des caractéristiques spatio-temporelles. Les limites de ces méthodes se placent au niveau de la complexité et / ou du temps de calcul. À titre d'exemple, entraîner des réseaux de neurones profonds $3D$ peut s'avérer difficile car ils requièrent une grande masse de données afin de fixer tous les paramètres de ces systèmes. Néanmoins peu de bases contiennent une grande quantité de données labellisées. Par ailleurs, étiqueter des STI est une tâche complexe car nous pouvons étudier des phénomènes ponctuels ou des comportements temporels. L'aspect $3D$ peut aussi rendre difficile la compréhension et l'explication du raisonnement logique menant à une décision particulière.

Dans ce contexte, nos motivations principales dans le cadre de ces travaux sont la prise en considération conjointement des deux domaines spatial et temporel pendant l'analyse des données pour que la nature des caractéristiques soit spatio-temporelle et ne pas la limiter à l'une des deux dimensions.

1.5 Objectif et contributions

Notre objectif réside ainsi dans l'analyse des STI, tout en considérant conjointement les domaines spatial et temporel. Cela doit permettre d'améliorer la compréhension de la dynamique de la scène observée ou d'un objet d'intérêt représenté par la STI. Les contributions de cette thèse se focalisent sur la proposition de deux approches principales pour analyser les STI, impliquées dans deux contextes applicatifs différents. Le point commun entre ces deux approches est le changement de représentations initiales des données dans le but d'extraire des caractéristiques spatio-temporelles natives.

La première contribution consiste à définir des caractéristiques permettant d'analyser l'évolution de la STI dans le temps. À titre d'exemple, considérons l'évolution des zones urbaines dans le temps, en télédétection, ou la reconnaissance d'action dans les vidéos. La complémentarité de l'évolution, qui est la stabilité, contient différentes informations particulières permettant la caractérisation de certains phénomènes. Par exemple, les zones urbaines ne changent pas significativement dans le temps (sauf dans le cas d'une catastrophe naturelle). Pour cela, notre premier axe de recherche se focalise donc sur la définition de caractéristiques qui mesurent la stabilité temporelle. La méthode est ensuite étendue / généralisée afin qu'elle puisse prendre en considération l'information spatiale.

La deuxième contribution méthodologique repose sur la représentation des données car elle influe sur la complexité des méthodes et aussi sur la qualité de leurs résultats. Travailler sur cette représentation s'avère pertinent afin de simplifier une STI tout en conservant les informations spatiales et temporelles des données. Le but principal est de passer des données $2D + t$ vers des données $2D$ tout en limitant la perte d'information. Cela produit des représentations planaires qui permettent aux méthodes d'apprentissage automatique profondes, conçues pour analyser les images $2D$, d'apprendre directement des caractéristiques spatio-temporelles. Cela permet de réduire les temps de calcul et d'augmenter le nombre de données annotées qui sont souvent difficilement accessibles dans le cadre des STI. La troisième contribution, se focalise sur l'explicabilité des différentes décisions prises par les réseaux de neurones car ils sont souvent considérés comme des boîtes noires non-interprétables. Cette dernière est liée à notre deuxième contribution. Cette explicabilité est basée sur des travaux existants qui sont ensuite appliqués sur les nouvelles représentations des données que nous proposons.

Un de nos objectifs est aussi le développement de méthodes génériques qui peuvent s'appliquer à différents domaines. Dans notre cas, nous avons choisi deux applications. La

première se focalise sur l'analyse des $STIS$ et la deuxième consiste en l'analyse de vidéos. Chacune de nos contributions a été testée et évaluée sur ces deux applications différentes. Dans la première, deux tâches spécifiques sont étudiées qui sont : (1) l'analyse de la couverture urbaine ; (2) et la classification de parcelles agricoles. La deuxième consiste en la classification de vidéos en violentes ou non-violentes.

Ce manuscrit de thèse est structuré en deux parties principales. La première partie entame une présentation d'un état de l'art sur les méthodes d'analyse de STI dans le chapitre 2. Ensuite le chapitre 3 introduit les applications ciblées. La deuxième partie contient deux chapitres où chacun présente une contribution méthodologique que nous proposons. Le chapitre 4 décrit la méthode qui mesure la stabilité temporelle tandis que le chapitre 5 expose la méthode de la création des représentations planaires. Les résultats obtenus sont présentés et commentés à la fin de chaque chapitre. Enfin, une conclusion générale avec les perspectives de recherche envisagées sont présentées.

Chapitre 2

Méthodes d'analyse de séquences temporelles d'images

La volonté trouve, la liberté choisit. Trouver et choisir, c'est penser.

– Victor Hugo, 1942

Les méthodes de l'état-de-l'art pour l'analyse des séquences temporelles d'images procèdent de différentes manières. Dans ce chapitre, nous présentons et discutons un diaporama de ces méthodes conçues pour traiter ce type de données. La section 2.1 introduit le

chapitre en discutant deux catégories de caractéristiques. Les méthodes d'extraction de caractéristiques présentées sont divisées en trois groupes. La section 2.2 présente le premier groupe des méthodes qui traitent les pixels temporels de façon individuelle. La section 2.3 expose le deuxième groupe des méthodes qui incluent de l'information spatiale aux pixels temporels. La section 2.4 est dédiée au dernier groupe qui englobe les méthodes spatio-temporelles. Une discussion est présentée dans la section 2.5.

2.1 Introduction

L'analyse automatique des séquences temporelles d'images permet d'automatiser des tâches visant à extraire des informations / connaissances à partir des données. De telles tâches étaient auparavant réalisées à la main par un opérateur humain. Toutefois, ces tâches étaient à la fois fastidieuses et aussi très limitées. À cela s'ajoute le risque d'erreur pouvant se répercuter dans l'écriture des règles expertes du logiciel. Les méthodes automatiques de traitement d'images permettent de s'affranchir de ces limites tout en étant plus précises [1]. Ces méthodes sont conçues pour résoudre différents problèmes, tels que la classification, la segmentation ou la détection d'objets. À l'instar des experts, les méthodes basées sur l'apprentissage sont capables d'apprendre ou d'optimiser leurs propres caractéristiques pour résoudre la tâche qui leur incombe. Dans la suite, nous nous sommes focalisés uniquement sur les méthodes d'extraction de caractéristiques adaptées aux problèmes de classification. Les méthodes considérées peuvent être rassemblées en deux catégories. La première catégorie est basée sur la définition de caractéristiques expertes, dites en anglais caractéristiques « *hand-crafted* ». La deuxième catégorie est basée sur les méthodes d'apprentissage automatique. Cet apprentissage peut se faire sur les caractéristiques « *hand-crafted* » ou directement sur les données brutes pour apprendre ces caractéristiques de façon automatique. Nous rappelons que le point commun entre ces deux catégories de méthodes est le changement de représentation des données initiales.

L'entraînement des méthodes d'apprentissage a pour but de trouver un certain nombre de paramètres automatiquement, conduisant à l'extraction des caractéristiques de haut niveau produisant des résultats optimisés pour un problème donné. Ces méthodes sont traditionnellement divisées en plusieurs catégories : supervisées, semi-supervisées et non-supervisées. Dans notre étude, nous nous concentrerons seulement sur l'aspect « caractéristiques » des méthodes de classification.

Toutefois, la qualité des résultats dépend de la nature et de l'aspect des caractéristiques extraites, qui elles-mêmes dépendent du point de vue adopté par la méthode. Dans notre cas, nous groupons les méthodes d'analyse des STI de la littérature en trois catégories. Le regroupement est basé sur la nature des caractéristiques extraites (*i.e.*, dimension traitée). Trois catégories de méthodes sont présentées :

1. Actuellement les méthodes neuronales offrent des performances supérieures à celles des humains sur des benchmarks tels que ImageNet [51].

1. la première catégorie de méthodes se concentre plutôt sur l'aspect temporel des données. Cette catégorie considère une STI comme un ensemble de pixels indépendants caractérisés par leur série temporelle conduisant à des séries temporelles $1D$ qui sont traitées individuellement ;
2. la deuxième catégorie de méthodes ajoute une information spatiale à l'information temporelle des pixels. La STI est traitée souvent image par image. Différentes caractéristiques spatiales et temporelles sont ensuite extraites puis agrégées afin de conduire à une décision globale. Il existe des méthodes, dites hybrides, qui traitent les deux aspects, spatial et temporel, de la STI séparément. Ensuite, une fusion des informations extraites est opérée au niveau décisionnel ;
3. la troisième catégorie représente les approches qui prennent directement en compte des caractéristiques spatio-temporelles calculées à partir du cube de données, par exemple, les caractéristiques convolutionnelles obtenues à partir d'un réseau de neurones profond $3D$.

Dans la suite de ce chapitre, nous détaillons ces trois catégories.

2.2 Les pixels comme des séries temporelles

Les méthodes d'analyse des séries temporelles d'images sont variées. Globalement, leurs processus de traitement considèrent la STI comme un ensemble de pixels temporels indépendants. Les méthodes les plus naïves considèrent que chaque pixel de la série est caractérisé par ses informations colorimétriques (*i.e.*, RGB). D'autres informations peuvent être calculées comme par exemple dans le cas des images satellitaires, des indices radiométriques calculés à partir de combinaisons des différentes bandes spectrales de l'image, tels que l'indice de végétation $NDVI$. Dans le cas des vidéos, il est possible de représenter les pixels dans un autre espace colorimétrique comme l'espace couleur HSV. Ensuite une méthode d'apprentissage machine est utilisée pour étiqueter les séries temporelles. Cette méthode peut être soit supervisée ou non-supervisée. La figure 2.1 illustre la chaîne de traitement générale des méthodes qui traitent les séries temporelles.

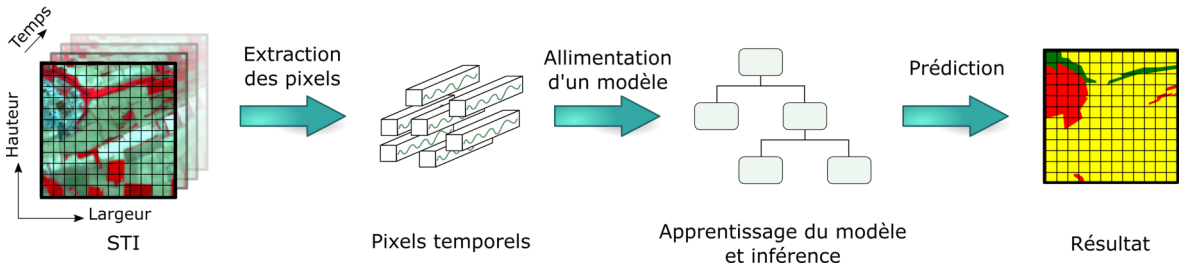


Figure 2.1 – Chaîne de traitement globale des méthodes d'analyse des séries temporelles.

Durant les années 90, des recherches ont été menées par la communauté de télédétection pour analyser les séries temporelles d'images [25, 87]. Certaines méthodes proposées

consistent à analyser, par exemple, l'occupation des sols observés depuis l'espace. Les séries temporelles ont permis d'étudier l'occupation des sols selon leurs types de changements qui peuvent être brusques ou observés sur le long terme. Les changements brusques sont généralement dus à des événements instantanés ou inattendus (*e.g.*, feu de forêt, tremblement de terre ou récolte sur une parcelle) car une discontinuité des valeurs dans le temps peut être remarquée. Les changements observés sur le long terme correspondent à des événements progressifs s'étalant dans le temps (*e.g.*, croissance de la végétation, urbanisation). En analyse de vidéo, le suivi d'objets est une problématique importante. Certaines méthodes résolvent ce problème en trouvant l'arrière plan de la scène. Ce dernier peut être modélisé par exemple par la valeur médiane d'un pixel temporel [87]. Dans la littérature, une séparation entre les méthodes d'analyse bi-dates et multi-dates est proposée car chacune a un avantage particulier par rapport à l'autre. Les méthodes bi-dates sont plus utiles pour l'analyse des changements brusques et les méthodes multi-dates sont plus utilisées pour analyser les changements observés sur le long terme. Dans la suite, ces deux groupes de méthodes sont abordés dans un premier temps. Dans un second temps, les méthodes basées sur un changement de représentation et les méthodes basées sur un apprentissage sont présentées.

2.2.1 Analyse des séries temporelles par paires

Les méthodes d'analyse par paires se basent sur l'analyse des valeurs des pixels à la même position à l'instant t et $t+1$, via de simples opérations, telles que la différence entre images [15] ou le rapport entre elles [62]. La localisation du changement est obtenue en appliquant un seuillage ou en utilisant un classificateur. Une autre méthode identifie le type de changement en construisant un vecteur sur un espace multidimensionnel à partir des deux valeurs de pixels [65]. La norme et la direction de ce vecteur donnent les informations sur le changement. Ce type de méthodes a été étendu pour analyser les STI en répétant le processus pour tous les couples d'images successives. Puis, en fonction de l'objectif, différentes techniques peuvent être appliquées pour agréger les résultats pour tous les pixels de même position [74, 150].

Ce type de méthodes permet d'extraire une information temporelle partielle. Quand une série d'images est considérée, la répétition du processus entre les paires d'images présente certaines limites. Par exemple, pour une série composée de 4 images l'analyse peut se formaliser selon deux stratégies différentes représentées comme suit :

$$f(f(I_1, I_2), f(I_3, I_4))$$
$$\text{ou } f(f(f(I_1, I_2), I_3), I_4)$$

où I_t est une image à l'instant t et f est une fonction mathématique d'analyse appliquée entre les paires d'images. Le résultat de f est une image. Les inconvénients de cette étude se présentent dans : (1) l'exploitation de l'information de la transition entre la paire (I_2, I_3) ; (2) l'ordonnancement des valeurs des pixels dans le temps affecte les résultats pour chacune des paires et les propriétés mathématiques de la fonction f appliquée (*e.g.*, associativité) [101].

Autrement dit, l'analyse des STI avec les méthodes par paires reste locale et la décision globale est faite en combinant les résultats obtenus entre chaque pair.

2.2.2 Le temps organise la série

Nous distinguons par la suite les méthodes qui prennent en compte conjointement toutes les images de la série temporelle. Ces méthodes considèrent les pixels comme un ensemble de mesures colorimétriques ordonnées chronologiquement, noté pixel temporel ou série temporelle de pixels $(p_t)_{t=1}^{T}$. En télédétection, nous trouvons des méthodes qui sont basées sur une approche de classification multi-date, comme l'analyse des trajectoires radiométriques [136] pour explorer l'évolution de la couverture de l'occupation des sols à travers le temps (*e.g.*, évolution de la végétation, changement saisonnier [119]). Toutefois, d'autres méthodes sont dédiées à l'analyse de l'évolution temporelle des données plus générales [109]. Ces dernières ont été expérimentées sur 47 bases de séries temporelles proposées par *University of California*[2]. Nous trouvons aussi une autre catégorie de méthodes qui analysent les séries temporelles de type satellitaires ou des signaux de type ECG[3] par l'exploitation des motifs d'évolution fréquents [85, 75]. Les caractéristiques de ces méthodes reposent sur des motifs qui sont majoritairement représentés dans les données. Par exemple, le motif « végétation $\rightarrow$ sol nu $\rightarrow$ urbain » illustre le phénomène d'urbanisation du terrain. Ce type de méthodes est basée sur une discrétisation des valeurs de la série en utilisant des approches symboliques.

Les méthodes basées sur des modèles statistiques sont aussi utilisées pour modéliser une observation qui évolue dans le temps. Les plus connues sont les modèles de Markov cachés, en anglais *Hidden Markov Model* (HMM), qui sont des modèles génératifs probabilistes de paramètres inconnus où les transitions du modèle seraient cachées pour l'utilisateur. Les HMM caractérisent les séries temporelles par une suite de transitions entre les états cachés du modèle. Les HMM ont été utilisés avec succès dans de multiples applications telles que la reconnaissance de la parole [105, 9], la biologie pour analyser des séries d'ADN [14] ou pour la modélisation de l'écriture manuscrite en temps réel [53, 86]. Cependant, les limites principales des HMM résident dans leurs comportements ayant un nombre d'états fini alors que leurs structures changent d'une application à une autre.

La classification des pixels temporels des STI peut se faire avec les approches d'apprentissage machine classiques, qu'elles soient supervisées ou non-supervisées. Certaines reposent sur des mesures comme la distance Euclidienne ou la *Dynamic Time Wrapping* (DTW) [99]. Par exemple, l'algorithme du plus proche voisin couplé à l'une de ces mesures peut être utilisé pour attribuer l'étiquette de la classe la plus similaire lors de l'analyse des $STIS$ [99, 100].

2. https://timeseriesclassification.com/dataset.php

3. ECG : est l'acronyme de **E**lectro**C**ardio**G**ramme qui est un examen consistant à mesurer l'activité électrique du cœur.

2.2.3 Transformation de l'espace de représentation

Les méthodes qui vont être présentées dans cette section se basent sur une transformation qui va projeter les données dans un nouvel espace de représentation avant de passer à l'étape d'analyse. L'objectif est de mettre en évidence des informations caractérisantes des données abstraites dans le domaine original. Le domaine le plus connu est le domaine fréquentiel. La transformée de Fourier permet de considérer les données dans cet espace fréquentiel et permet ainsi d'étudier les spectres des différentes séries temporelles [4]. Une telle transformation nécessite un échantillonnage temporel régulier. Il y a aussi la transformée en ondelette qui s'affranchit des limites d'échantillonnages temporels réguliers [19]. De telles transformations permettent d'analyser l'auto-corrélation et la corrélation croisée entre les séries temporelles. Le problème principal de ces transformations est qu'elles ont besoin de séries temporelles très longues ce qui n'est pas forcément le cas dans certaines applications.

D'autres méthodes procèdent différemment. L'espace induit par des caractéristiques bien choisies est plus discriminant que les données brutes pour la tâche de classification. Parmi ces méthodes, nous citons l'adaptation des caractéristiques *Scale Invariant Features Transform* ($SIFT$) au niveau temporel [10]. Ensuite, les pixels temporels sont caractérisés par un sac de mots de $SIFT$. Ce dernier est un histogramme qui représente les fréquences d'apparition des $SIFT$ dans le pixel temporel. D'autres types de caractéristiques [23] adaptées pour traiter des séries temporelles peuvent être prises en compte. Les plus basiques sont des mesures statistiques telles que le minimum, le maximum, la moyenne, la médiane ou l'écart-type. D'autres sont un peu plus avancées comme l'asymétrie, l'aplatissement, l'énergie ou l'entropie. Les auteurs de [23] proposent la bibliothèque TSfresh [4] qui contient un panorama de caractéristiques dédiées aux séries temporelles. Ces caractéristiques peuvent être utilisées pour traiter les signaux biomédicaux (*e.g.*, EEG et ECG), les données financières (*e.g.*, taux boursiers) ou des séries générées par des appareils industriels (*e.g.*, images satellitaires, capteurs de gaz ou excitation laser).

2.2.4 Vers l'apprentissage des caractéristiques

Plus récemment, les réseaux de neurones profonds ont été considérés pour la classification des séries temporelles. Dans la littérature, ces méthodes sont regroupées dans deux grandes familles : les réseaux de neurones récurrents et les réseaux de neurones convolutifs. Dans la suite, nous utilisons les acronymes anglophones de ces deux types de familles qui sont respectivement RNN comme *Recurrent Neural Network* et CNN comme *Convolutional Neural Network*. Les RNN ont été conçus pour traiter les données séquentielles comme les séries temporelles en proposant de nouveaux modèles [12] dotés d'une mémoire qui leur permettent de traiter des séquences de longueurs différentes. Cependant, ils souffrent de certaines limites lors de l'apprentissage avec l'atténuation du gradient [52]. Pour cela, deux améliorations sont proposées qui sont les réseaux *Gated Recurrent Unit* (GRU) et les

4. https://tsfresh.readthedocs.io

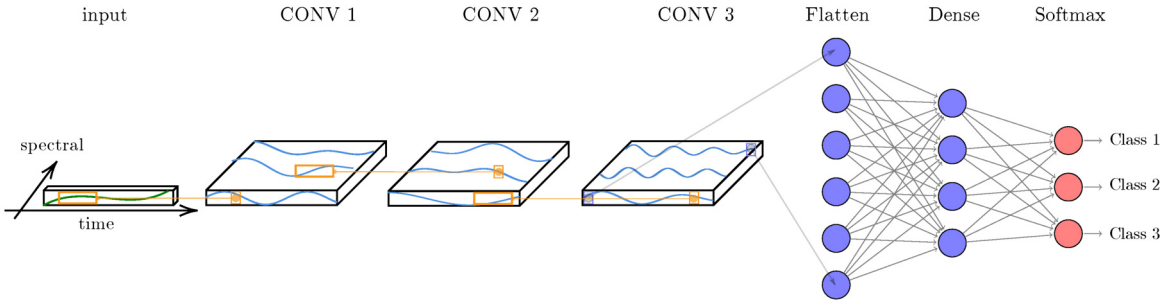


Figure 2.2 – Illustration de $TempCNN$ [98] (figure reprise de [98]).

Long Short-Term Memory ($LSTM$). Le plus populaire est le $LSTM$. La force des réseaux récurrents réside dans leur capacité de mémoire pour le traitement des séquences. Ce type de modèle a été utilisé avec succès dans différentes applications. Nous citons par exemple, la reconnaissance de la parole [46], la reconnaissance d'émotions [147] ou même la cartographie de l'occupation des sols dans le domaine de l'observation de la Terre [112, 59].

Au départ, les CNN ont été conçus pour analyser des images $2D$ et les convolutions ne s'appliquaient que sur le domaine spatial [55]. Au fil du temps, ils ont été adaptés aux séries temporelles. Par exemple, l'architecture $TempCNN$ a été proposée pour la classification des $STIS$ [98] en utilisant des convolutions $1D$ appliquées dans le domaine temporel. La figure 2.2 illustre l'architecture de $TempCNN$. Les architectures des CNN classiques, tels que $ResNet$, ont également été adaptées pour la classification des séries temporelles et expérimentées dans les bases proposées par *University of California*[5] [37]. Certaines études montrent que les méthodes par réseaux de neurones profonds permettent d'avoir de meilleurs résultats en terme de classification des séries temporelles que les approches classiques, telles que les forêts aléatoires [37]. Cela est dû aux caractéristiques extraites qui sont optimisées pour la tâche ciblée.

Une autre stratégie consiste à encoder les séries temporelles en des représentations $2D$. Cela permet l'utilisation des CNN classiques $2D$ pour la classification de ces représentations $2D$. Parmi ces méthodes, nous citons les *Recurrence Plots* (RP) [83] qui sont des représentations $2D$ permettant d'explorer les récurrences des trajectoires de chaque point dans l'espace des phases. Dans la théorie des systèmes dynamiques, l'espace des phases représente tous les états possibles d'un système. Soit le pixel temporel $p = (p_t)_{t=1}^{T}$ où T est la longueur de p. Les trajectoires pour chaque point de la série sont $\overrightarrow{p} = (\overrightarrow{p_1} : (p_1, p_2) \ldots, \overrightarrow{p_{T-1}} : (p_{T-1}, p_T))$ avec $T-1$ états. La génération de la matrice des RP est faite par le calcul de la distance entre tous les états des trajectoires comme indiqué dans l'équation 2.1 :

$$RP(i,j) = \begin{cases} 1, & \text{si } ||\overrightarrow{p_i} - \overrightarrow{p_j}|| \leq \varepsilon \\ 0, & \text{sinon} \end{cases} \quad \text{avec } i,j \in \big([1, T-1]\big)^2 \tag{2.1}$$

5. https://timeseriesclassification.com/dataset.php

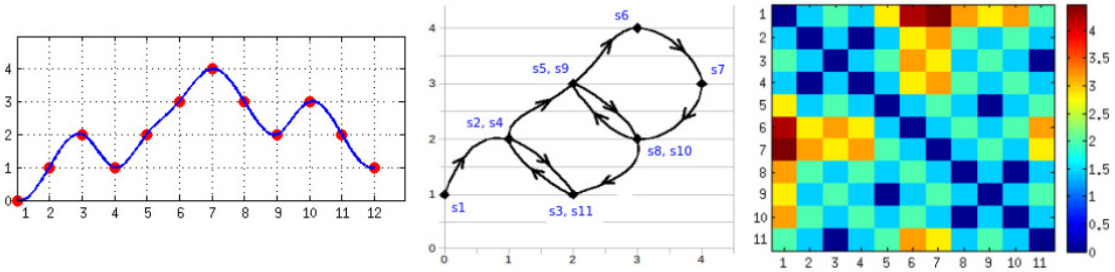


Figure 2.3 – Transformation d'une séquence temporelle en une matrice de *Recurrence Plot*. (Gauche) la séquence temporelle de longueur 12. (Milieu) plan 2D de l'espace des trajectoires. (Droite) la matrice résultante de *Recurrence Plot* de taille 11×11 [50].

La figure 2.3 illustre les étapes de la transformation d'une séquence temporelle en une matrice de RP. Le seuil ε peut être calculé de deux manières différentes. La première façon est basée sur la mesure de la densité des points selon un pourcentage. La deuxième façon est aussi basée sur un pourcentage mais de la distance maximale entre les points. Le choix du calcul du seuil ε dépend fortement de la nature des séquences étudiées.

Une autre méthode qui permet d'avoir une représentation $2D$ de la corrélation temporelle entre les points d'une séquence est la *Gramian Angular Field* [142]. La première étape de cette méthode consiste à normaliser les données entre $[-1, 1]$. Ensuite les coordonnées polaires qui sont le rayon r et l'angle θ sont calculées pour chaque point du pixel temporel p. Puis deux types de corrélations temporelles entre les points peuvent être extraites. La première est le cosinus de la somme des angles qui est la *Gramian Angular Summation Field* ($GASF$) et la deuxième est le sinus de la différence des angles qui est la *Gramian Angular Difference Field* ($GADF$). L'équation 2.2 présente respectivement les deux formules pour calculer la $GASF$ et la $GADF$:

$$\begin{aligned} GASF(i,j) &= cos(\theta_i + \theta_j),\ \forall i,j \in [1,T] \\ GADF(i,j) &= sin(\theta_i - \theta_j),\ \forall i,j \in [1,T] \end{aligned} \tag{2.2}$$

Enfin, il existe les méthodes qui utilisent la transformée de Fourier à temps court, dite en anglais *Short Time Fourier Transform* ($STFT$), pour générer plusieurs spectres avec différentes fréquences conduisant à créer un spectrogramme [93]. La figure 2.4 illustre les résultats visuels des RP non binarisés, $GASF$, $GADF$ et $STFT$ sur deux séries temporelles issues de la base *GunPoint*. Ces stratégies permettent de bénéficier des avantages des CNN $2D$ qui peuvent déjà être entraînés à diverses tâches de vision par ordinateur, par exemple la classification d'images naturelles où le modèle est entraîné sur la base ImageNet [110] qui contient des millions d'exemples et ce dans 1000 catégories différentes.

Bien que ces approches basées sur les réseaux de neurones profonds offrent des résultats prometteurs, elles restent tout de même limitées dans le cas d'analyse des STI car la dimension spatiale des images n'est pas prise en compte et les pixels temporels sont traités de manière indépendante. Cependant, dans certaines applications, il est crucial de prendre

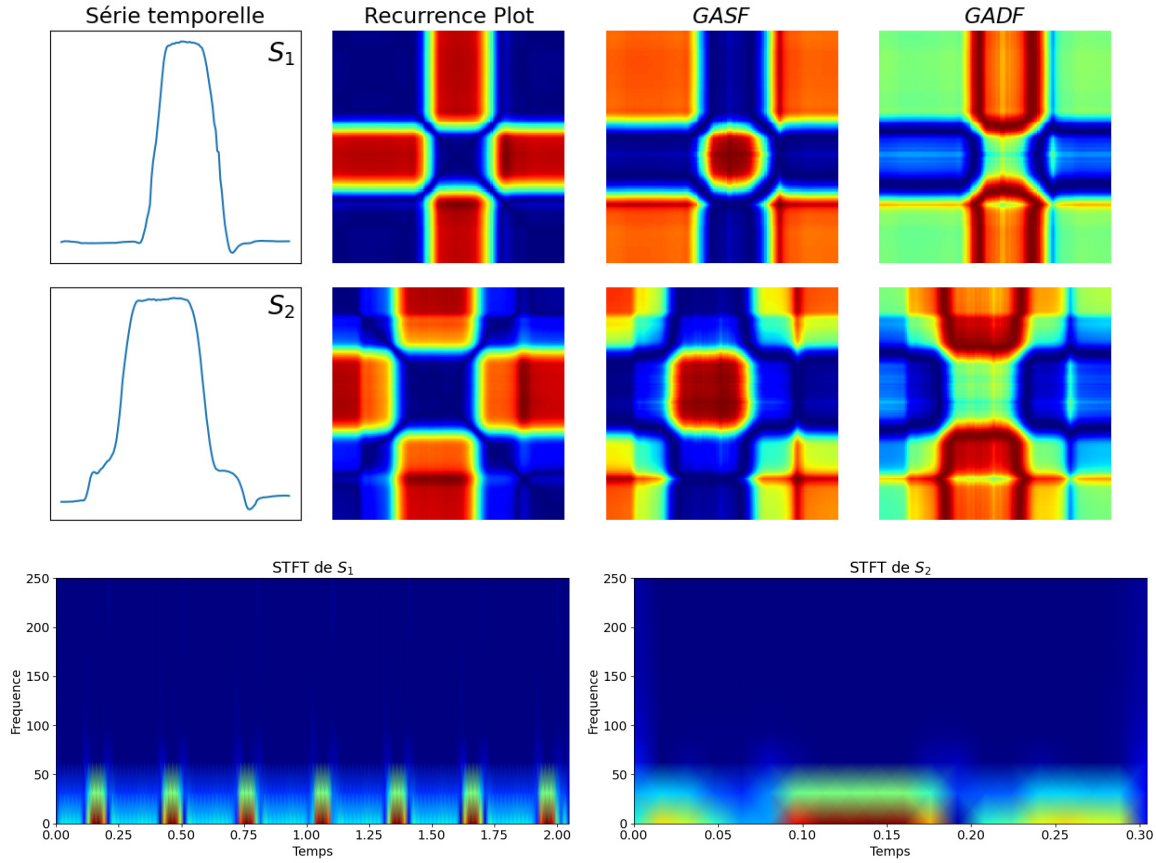


Figure 2.4 – Illustration des résultats visuels des RP, $GASF$, $GADF$ et $STFT$ sur deux séries temporelles issues de la base *GunPoint*.

également en compte l'aspect spatial des données $2D + t$ pour discriminer avec précision des classes complexes comme c'est le cas dans le cadre de nos domaines applicatifs.

2.3 Ajout d'information spatiale aux pixels temporels

Les méthodes temporelles se limitent aux informations colorimétriques ou spectrales collectées à chaque date. L'inclusion d'information spatiale lors de l'analyse des STI est importante afin de s'affranchir des limites des méthodes temporelles. L'idée principale d'inclusion des informations spatiales est de prendre en compte l'information collectée en fonction du voisinage spatial des pixels dans le domaine de l'image. Les méthodes permettant l'intégration de telles informations peuvent être groupées en trois familles. La première famille traite la STI, image par image, pour extraire des caractéristiques spatiales qui évoluent dans le temps et qui sont par la suite considérées comme un sac de mots. La deuxième famille vise à rajouter des caractéristiques spatiales aux séries temporelles. La dernière famille procède par régions.

2.3.1 Méthodes basées sur les sacs de mots

Les méthodes de cette famille traitent la STI de deux manières différentes. La première stratégie consiste, pour chaque image, à calculer différentes caractéristiques spatiales et la deuxième vise à calculer des caractéristiques temporelles. Ensuite, un vocabulaire de mots est construit en regroupant toutes les caractéristiques récoltées en K groupes. Les centroïdes des groupes représentent le vocabulaire des mots, c'est-à-dire K mots. Après, chaque STI sera caractérisée par un sac de mots, noté BoW comme *Bag of Words*. Ce dernier est un histogramme de fréquence de longueur K contenant le nombre d'occurrences des mots présents dans le vocabulaire de mots d'une STI. Ce processus est appliqué sur toutes les STI d'une base d'apprentissage. Au moment du test, la nouvelle STI est représentée avec un BoW et comparée avec les BoW de la base d'apprentissage. La classification peut se faire avec, par exemple un SVM. La figure 2.5 illustre le processus de la création du vocabulaire de mots et la représentation d'une STI en BoW.

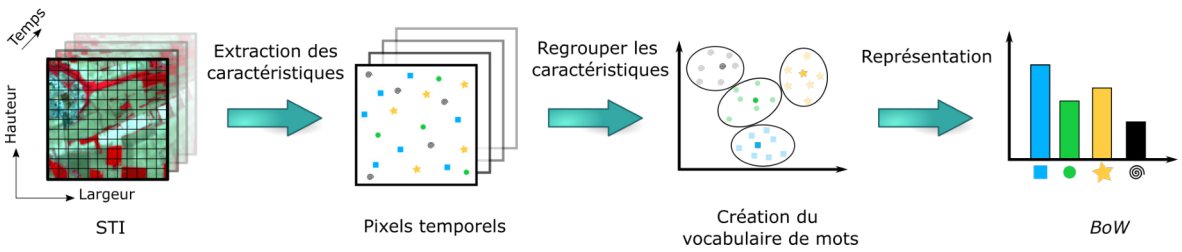


Figure 2.5 – Processus global de la création d'un vocabulaire de mots et la représentation de la STI en BoW.

Ce type de méthodes a été utilisé avec succès pour la reconnaissance d'actions à partir de vidéos. Parmi les caractéristiques utilisées, nous citons les histogrammes des gradients orientés [56], notés HOG qui est l'abréviation du terme anglais *Histogram of Oriented Gradient*. Ce dernier représente l'information de l'orientation du gradient dans des petits patchs de l'image sous forme d'histogramme. Au départ, les composantes du gradient G_x et G_y sont calculées sur les deux directions du plan de l'image. Elles peuvent être calculées par convolution avec deux filtres F_x et F_y appliqués sur l'image. L'étape suivante consiste à calculer la norme et la direction du gradient. Enfin, le BoW est construit en comptant le nombre d'apparitions des normes associées à leurs directions [117, 27].

Une autre méthode de reconnaissance d'action nommée HOG^2 est proposée dans [95]. Dans cette dernière, des HOG sont calculés pour toutes les images de la STI. Ensuite, tous les descripteurs HOG sont agrégés pour construire une image $2D$ sur laquelle d'autres HOG sont calculés afin qu'ils soient spatio-temporels. Enfin, un BoW est calculé sur l'image obtenue avec HOG^2. Il existe un autre type de caractéristiques basé sur la détection de points d'intérêt dans l'image. Ce dernier est la transformation de caractéristiques visuelles invariantes à l'échelle, plus connue sous l'acronyme $SIFT$ [130] qui est « *Scale-Invariant Feature Transform* ». Cette caractéristique représente les informations locales d'une image avec une invariance : à l'échelle, au cadrage, à l'angle de la prise d'image

et aussi à la luminosité. Les BoW des $SIFT$ ont aussi été appliqués avec succès dans l'analyse des vidéos [152].

2.3.2 Méthodes d'inclusion de l'information spatiale

Différentes recherches ont mené à la proposition de stratégies permettant d'intégrer des informations spatiales aux pixels temporels. Certaines méthodes commencent par segmenter les images de la STI individuellement en régions [100]. Ensuite, chaque pixel temporel est enrichi avec des caractéristiques spatiales calculées au niveau de la région à laquelle il appartient et ce pour chaque image de la série. Ces caractéristiques peuvent être colorimétriques, morphologiques, géométriques [100, 26]. La méthode baML [84] utilise un $CNN\ 1D$ appliqué sur le domaine temporel uniquement. Puis différentes caractéristiques spatiales sont calculées au niveau de chaque date de la série et ce pour chaque pixel. Ces caractéristiques spatiales peuvent être la moyenne ou l'écart-type des valeurs des pixels autour du pixel concerné. Les auteurs de la méthode ont également rajouté les coordonnées (x, y) dans le but de lisser les résultats. Le processus d'enrichissement des pixels temporels avec l'information spatiale est schématisé dans la figure 2.6.

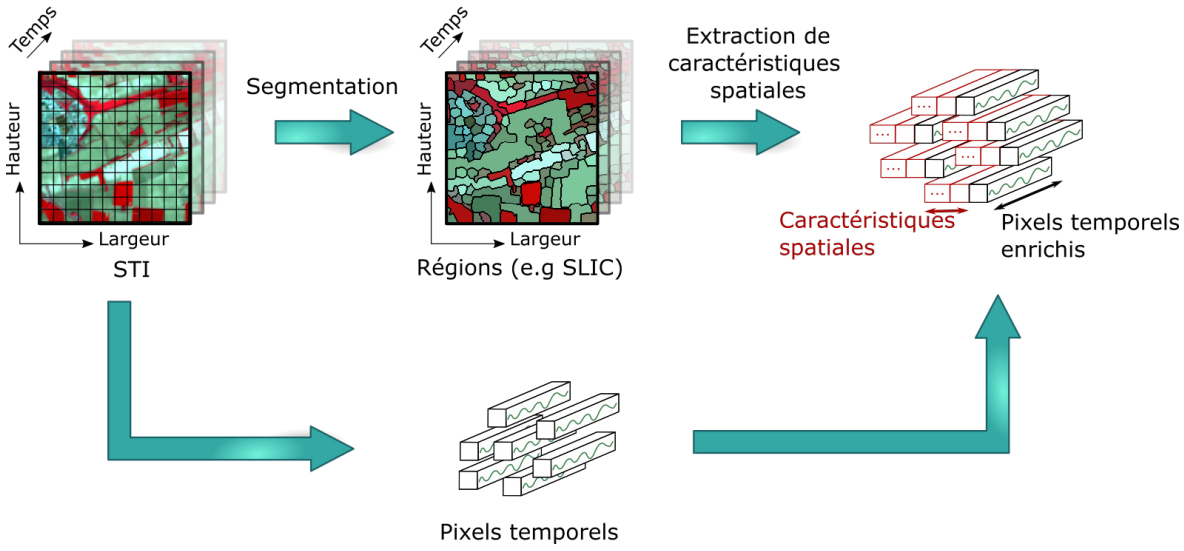


Figure 2.6 – Enrichissement des pixels temporels avec de l'information spatiale.

La segmentation des images en régions est une étape indispensable pour ces méthodes. Ces dernières sont diverses et leur choix est crucial car il influe sur la qualité des régions obtenues à partir desquelles les caractéristiques spatiales sont calculées. Parmi ces algorithmes, nous citons l'algorithme de *Felzenszwalb-Huttenlocher* (FH) [88]. Ce dernier considère les pixels de l'image comme les sommets d'un graphe et les régions sont construites en calculant un arbre couvrant de poids minimal. Nous pouvons également citer les approches qui se basent sur un regroupement non-supervisé des pixels : la *Simple Linear Iterative Clustering* ($SLIC$) [1] et le *MeanShift* [24]. Ces dernières projettent les pixels dans un espace à

cinq dimensions (les couleurs RGB et les coordonnées (x, y) des pixels). Les régions sont constituées grâce à un algorithme de *clustering* tel que K-Moyennes. Certains chercheurs ont expérimenté différents algorithmes de segmentation sur des images aériennes [8]. Dans le cadre des STI, *MeanShift* a aussi été utilisé avec succès [100].

Les opérateurs de morphologie mathématique ont été utilisés pour enrichir les pixels temporels. Parmi les recherches menées dans ce contexte, les profils d'attributs morphologiques et les pyramides de caractéristiques spatiales basées sur les pixels sont employés généralement pour des tâches de détection de changement en télédétection [34, 134]. Dans le domaine du suivi d'objets, une approche repose sur les HOG qui ont été considérés pour capturer des informations contextuelles pour la détection et le suivi des piétons dans les vidéos [11]. Chaque image peut également être traitée à l'aide d'un CNN $2D$ pour extraire des caractéristiques convolutionnelles qui sont empilées et passées à un classificateur pour fournir une classification globale de la STI [133].

2.3.3 Méthodes basées régions

Une autre catégorie de méthodes traite directement les régions obtenues par les algorithmes de segmentation. Cette catégorie de méthodes caractérise chaque région par un ensemble de caractéristiques calculées sur les pixels qui lui sont associés. Par exemple, les différentes régions obtenues par une segmentation sont analysées en utilisant des graphes d'évolution temporelle [69]. Cette méthode commence par segmenter l'ensemble

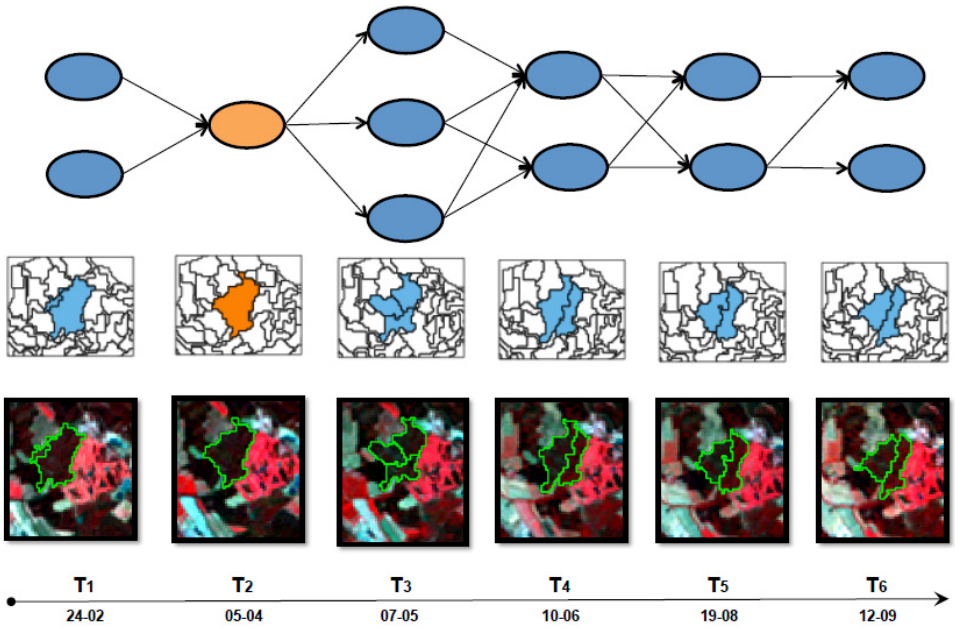


Figure 2.7 – Graphe d'évolution identifié par l'objet de référence en orange de la *Vallée du Libron* (figure tirée de [69]).

des images individuellement. Puis, les objets qui couvrent les zones les plus grandes sont sélectionnés comme des objets de référence. Ensuite, pour chaque objet de référence, un graphe d'évolution temporelle est construit. Enfin, un algorithme de regroupement est appliqué sur les graphes d'évolution. La figure 2.7 illustre un exemple de graphe d'évolution.

Dans le cadre de la classification de parcelles agricoles à partir de $STIS$, les auteurs de [42] ont récemment proposé une stratégie hybride s'appuyant sur les réseaux de neurones profonds. La première étape consiste à extraire des caractéristiques spatiales. Ils commencent par sélectionner aléatoirement un ensemble de pixels qui sont encodés avec un réseau inspiré des méthodes traitant les données de type nuages de points $3D$. Ce dernier permet d'apprendre des descripteurs statistiques de premier ordre de la distribution spectrale des observations. Ces caractéristiques sont ensuite combinées avec des caractéristiques temporelles extraites à l'aide d'un deuxième réseau de neurones basé sur une architecture de *self-attention* [6], pour produire finalement un résultat de classification.

Les approches présentées dans cette section permettent de combiner des caractéristiques spatiales avec des caractéristiques temporelles, mais ne conduisent pas à des caractéristiques nativement spatio-temporelles. Par exemple les méthodes des BoW sont basées sur un regroupement des valeurs. Cela fait perdre quelques informations en les limitant seulement à K qui représente le nombre de groupes. Quant à l'analyse par régions, les méthodes présentées limitent les caractéristiques à la région elle-même et ne permettent pas de traiter la transition entre les régions. Dans ce contexte, l'analyse des deux domaines spatial et temporel simultanément permet l'extraction de caractéristiques spatio-temporelles natives. De plus, certaines applications ont besoin d'un tel mode d'analyse afin d'avoir des caractéristiques globales de la STI.

2.4 Méthodes spatio-temporelles

Cette catégorie de méthodes est conçue pour produire des caractéristiques spatio-temporelles qui sont calculées en faisant intervenir simultanément les domaines spatial et temporel de la STI. La plupart des recherches menées sur ces caractéristiques ont été effectuées dans le contexte de l'analyse de vidéos telles que la reconnaissance d'actions, la détection et le suivi d'objets, le résumé d'une vidéo ou également l'analyse « en ligne » de l'écriture. Les premières méthodes proposent une adaptation des caractéristiques « *hand-crafted* » $2D$ vers le $2D + t$ pour les rendre spatio-temporelles. Ensuite, il y a les méthodes statistiques. Enfin, nous trouvons les méthodes récentes basées sur l'apprentissage automatique.

6. Les modèles dits *self-attention* sont basés sur l'attention cognitive afin de donner plus d'importance aux parties discriminantes des données d'entrée et d'estomper le reste.

2.4.1 Extension des caractéristiques *hand-crafted*

Certaines caractéristiques *hand-crafted* qui étaient initialement définies sur les images $2D$ ont ensuite été adaptées aux données $2D+t$. Parmi ces caractéristiques, nous trouvons celles de Haralick qui ont servi pour faire la segmentation d'images médicales de type tomodensitométrie [131] ou les motifs locaux binaires (LBP) pour la reconnaissance d'action [2]. Il y a aussi les histogrammes des flux [114], dits en anglais *Histograms Of Flow* (notés HOF), qui se basent sur le même principe que les HOG sauf que les données d'entrée sont les flux optiques de la STI. L'extension des histogrammes de gradient a donné naissance aux $3D$-HOG. Ces derniers ont été appliqués pour la détection et la classification des piétons et véhicules dans les routes urbaines [16]. Nous citons aussi les histogrammes de flux $3D$, notés $3D$-HOF, utilisés pour la détection des humains [30]. Certains travaux ont proposé une méthode qui combine les HOG et les HOF dans le but d'avoir une représentation de l'espace-temps pour une meilleure reconnaissance des actions [72].

Les $SIFT$ ont aussi été adaptés aux données STI, où le gradient temporel G_t est estimé par la composante temporelle du gradient. Ensuite, la norme et les deux orientations sont obtenues à partir des gradients calculés. Les $3D$-$SIFT$ ont été utilisés avec succès pour la reconnaissance des actions [115].

Dans une autre proposition [149], les mouvements des $SIFT$ ($MoSIFT$); le flux du mouvement est calculé seulement entre les points d'intérêt détectés par les $SIFT$. Une autre caractéristique qui est basée sur les détecteurs de coins de Harris [30] permet de détecter les *Spatial-Temporal Interest Point* ($STIP$) [47]. Les points d'intérêt sont localisés en utilisant une matrice Hessienne-Laplace H définie, pour un pixel p à la position (x, y) et à l'instant t par :

$$H(x,y,t) = \begin{pmatrix} \frac{\partial^2 p}{\partial x^2} & \frac{\partial^2 p}{\partial x \partial y} & \frac{\partial^2 p}{\partial x \partial t} \\ \frac{\partial^2 p}{\partial x \partial y} & \frac{\partial^2 p}{\partial y^2} & \frac{\partial^2 p}{\partial y \partial t} \\ \frac{\partial^2 p}{\partial x \partial t} & \frac{\partial^2 p}{\partial y \partial t} & \frac{\partial^2 p}{\partial t^2} \end{pmatrix} \tag{2.3}$$

D'autres auteurs considèrent le vocabulaire des points d'intérêt des $STIP$ comme un graphe orienté saillant [77]. Ce dernier contient des arêtes entre les points spatiaux mais aussi des arêtes entre les images de la STI afin d'avoir un graphe saillant spatio-temporel entre les points $STIP$. La figure 2.8 illustre la chaîne de traitements proposée dans [77].

Une autre catégorie de méthodes analyse des données de type nuages de points $3D$. Ces dernières ont été utilisées avec succès pour la reconnaissance d'actions [138, 89, 122]. Ce type de données permet d'avoir une information sur le changement géométrique tout en étant robuste aux vidéos mal contrastées. L'acquisition de ce type de données se fait avec des capteurs de profondeur afin d'avoir une représentation $3D$ du corps humain. Par

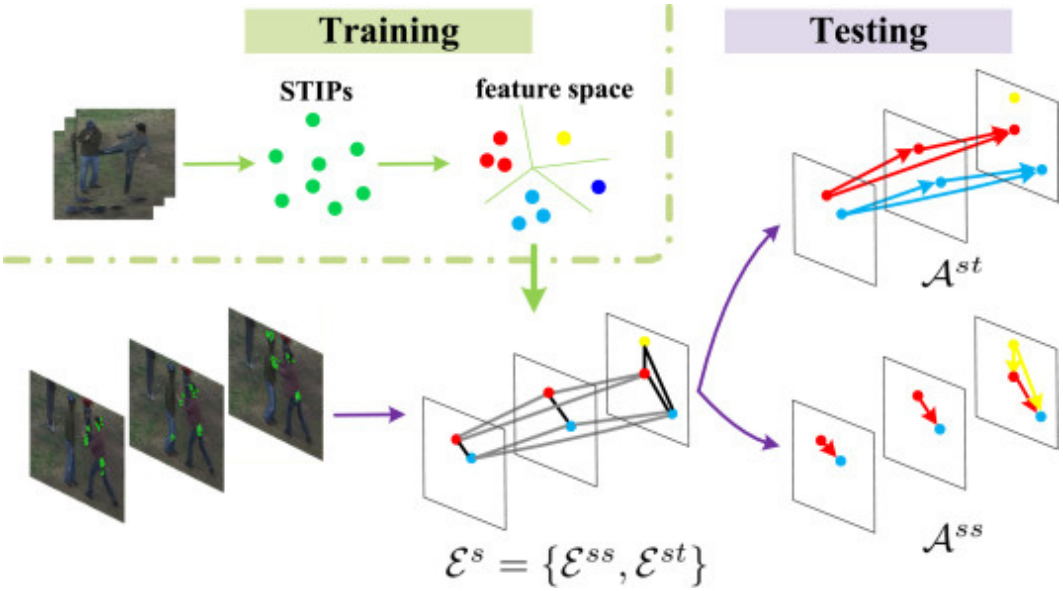


Figure 2.8 – Extraction et création du graphe spatio-temporel pour la reconnaissance d'action [77]. $\mathcal{E}^s$ est l'ensemble des arêtes du graphe saillant où $\mathcal{E}^{ss}$ et $\mathcal{E}^{st}$ représentent respectivement les arêtes spatiales et temporelles. $\mathcal{A}^{st}$ et $\mathcal{A}^{ss}$ sont des graphes générés à partir des arrêtes $\mathcal{E}^{st}$ et $\mathcal{E}^{ss}$.

la suite, le squelette de celui-ci est représenté par des nuages de points au cours du temps. L'exploitation de l'espace géométrique permet le calcul d'autres caractéristiques telles que l'azimut pour avoir une information de la direction du déplacement à chaque instant [138, 89]. Une autre méthode propose de représenter chaque squelette en utilisant les rotations 3D relatives entre les différentes parties du corps puis analyse les mouvements subis dans le temps [135]. Les auteurs de [76] ont proposé d'abord d'extraire des caractéristiques squelettiques (*e.g.*, position de chaque point, couleur RGB) qui sont par la suite classifiées de façon supervisée.

Les caractéristiques *hand-crafted* ont été utilisées avec succès lors de l'analyse des STI en considérant les deux domaines spatial et temporel simultanément. Toutefois, ces dernières ne sont pas optimisées par rapport à un objectif mais elle peuvent être utilisées pour résoudre plusieurs problèmes. De plus, ces caractéristiques ne généralisent pas certaines représentations et surtout dans le cas de la reconnaissance des actions.

2.4.2 Méthodes statistiques

Les méthodes statistiques sont aussi utilisées pour généraliser différentes observations qui sont supposées suivre un processus stochastique. Certains chercheurs ont modélisé grâce à des HMM diverses applications de reconnaissance de gestes. Par exemple, les auteurs de [146] utilisent un HMM qui traite la vélocité des objets caractérisés par un ensemble de caractéristiques comme les coordonnées du barycentre (x, y) et l'angle de dé-

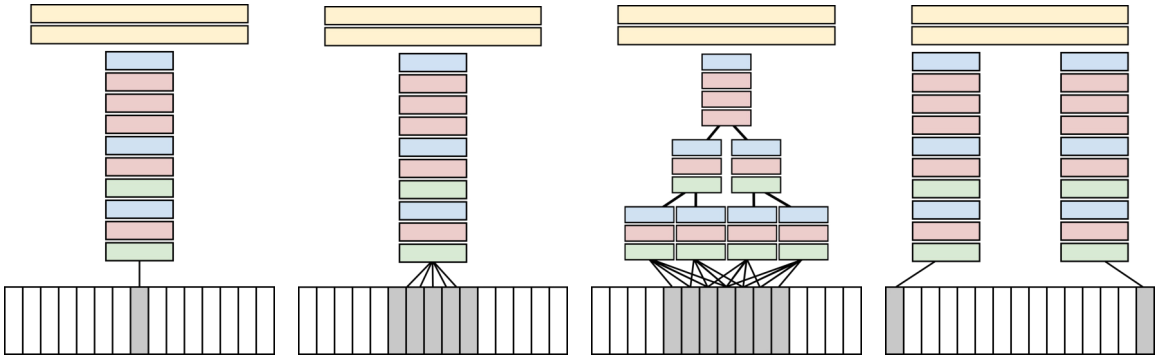
placement. Par contre, l'ensemble des états des gestes étudiés reste limité à cinq seulement. Un autre type d'applications est la vérification de l'authenticité des signatures en temps réel [53]. La base de cette méthode est l'enregistrement de plusieurs informations comme la vitesse et l'accélération du stylo sur la feuille et les coordonnées. Finalement, le modèle HMM utilise le chemin de Viterbi qui servira par la suite à vérifier l'authenticité de la signature [80].

Nous trouvons également dans la littérature les champs aléatoires conditionnels, notés CRF comme *Conditional Random Field* en anglais, qui sont aussi utilisés pour la reconnaissance de mouvement [123]. Pour ce faire, chaque image d'une vidéo est représentée par un ensemble de caractéristiques de forme, de contour et d'angle de trajectoire. Ensuite, le CRF prédit l'action selon toutes les prédictions associées aux images de la vidéo. Ces modèles ne capturent pas des informations sur les états cachés mais ils fournissent une étiquette à chaque observation de la séquence. Une amélioration des CRF est le $HCRF$, comme *Hidden Conditional Random Field*. $HCRF$ permet d'avoir des informations sur les états cachés dans le temps. Les $HCRF$ sont directement utilisés pour faire la reconnaissance des gestes [54]. Dans une autre méthode, un SVM est couplé avec les $HCRF$ pour faire la reconnaissance d'activité [116]. Le SVM est utilisé sur les caractéristiques contextuelles et de mouvements pour avoir un vecteur de caractéristiques de haut niveau qui sera traité par le $HCRF$.

Les méthodes présentées dans cette section dépendent des caractéristiques utilisées pour pouvoir classifier les données. De plus, une bonne maîtrise de l'aspect probabiliste est indispensable. Enfin, le manque d'interprétabilité est une des faiblesses de cette catégorie de méthodes.

2.4.3 Caractéristiques reposant sur l'apprentissage profond

Récemment en intelligence artificielle, les approches basées sur les réseaux de neurones profonds ont été utilisées pour l'apprentissage de caractéristiques spatio-temporelles et leurs implications dans diverses tâches de vision par ordinateur [28, 6]. Certains auteurs ont comparé quatre architectures où les données d'entrée sont traitées différemment pour la classification de vidéos [67, 63]. Le premier modèle est un simple CNN, initialement conçu pour classer des images [110], où les convolutions ne sont appliquées que sur le domaine spatial. Il procède alors image par image pour traiter une vidéo. Les trois autres modèles se basent sur une fusion des caractéristiques extraites des STI qui peuvent être précoces, progressives ou tardives. La figure 2.9 illustre les quatre modèles. Le modèle de fusion précoce est basé sur des convolutions 3D appliquées sur T images pour extraire les informations spatio-temporelles. Le modèle à fusion progressive combine les stratégies de fusion des deux modèles précédents. Ce dernier utilise des convolutions $3D$ sur une fenêtre cubique qui avance progressivement dans le temps. Après chaque pas, une fusion des valeurs des caractéristiques est appliquée afin qu'elles s'adaptent au contenu de la scène. Cette fusion peut se faire par une addition ou multiplication terme à terme. Ce type de modèle généralise mieux mais il est lent et consomme beaucoup de ressources. Enfin le modèle de



(a) Une seule image (b) Fusion précoce (c) Fusion progressive (d) Fusion tardive

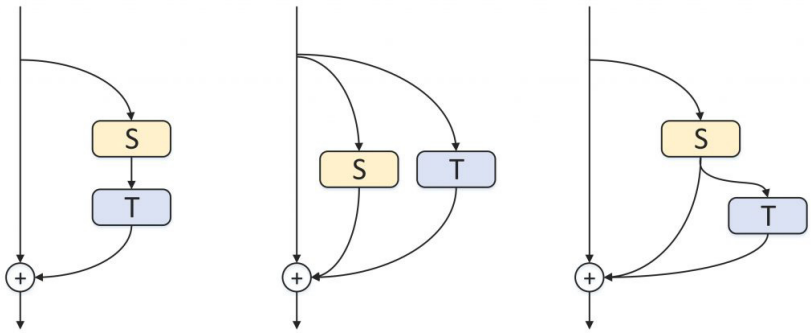
Figure 2.9 – Présentation des différents niveaux de fusion des caractéristiques pour la classification de vidéos [67].

fusion tardive est complètement similaire à un $CNN\ 2D$ classique, sauf que deux images sont fournies en entrée. Ces dernières sont traitées séparément par le CNN. Ensuite, une couche linéaire est appliquée sur la concaténation des caractéristiques des deux images afin d'apprendre le mouvement subi entre ces deux images.

Par la suite, plusieurs autres méthodes plus performantes ont été proposées où les différentes fusions s'appliquent sur des caractéristiques apprises sur d'autres types d'informations, telles que le flux optique. Nous allons les décrire par la suite en les regroupant selon les trois types de fusions présentés précédemment. Puis, nous évoquerons aussi des méthodes basées sur les nuages de points.

2.4.3.1 Fusion précoce

Cette catégorie de méthodes considère uniquement la vidéo RGB brute sans autre information supplémentaire. En 2015, le modèle $C3D$ pour *Convolutional* $3D$ [64] a été proposé. Les convolutions de $C3D$ traitent simultanément le domaine spatial et temporel. $C3D$ a été utilisé avec succès pour classer des vidéos d'action [64]. Les auteurs de [18] ont proposé *Inflated 3D* ($I3D$) où ils ont rajouté au modèle $C3D$ une couche de normalisation du lot (*batch*) après chaque couche de convolution. Un pas temporel de 1 est utilisé dans la première couche de pooling afin de pouvoir utiliser un lot de vidéos plus grand pendant l'entraînement. Nous trouvons aussi le modèle $P3D$ pour *pseudo* $3D$ [104]. $P3D$ s'inspire du modèle *ResNet* sauf que toutes les couches ont été adaptées aux données $3D$. $P3D$ rajoute des connections résiduelles entre les différentes couches du modèle. Ces connections résiduelles ont été étudiées de différentes manières. Les domaines spatial et temporel sont traités différemment afin que les caractéristiques spatio-temporelles soient les plus discriminantes possible. La figure 2.10 illustre les différentes stratégies des connections résiduelles étudiées par les auteurs de $P3D$ [104]. Dans [133], plusieurs formes de convolutions spatio-temporelles pour l'analyse de vidéos sont étudiées. Leurs conclusions montrent les avantages en terme de précision des $CNN\ 3D$ par rapport aux



(a) P3D-A (b) P3D-B (c) P3D-C

Figure 2.10 – Les différents traitements des connections résiduelles du modèle $P3D$ [104] : (a) traitement du domaine spatial suivi du temporel ; (b) traitement des domaines spatial et temporel séparément ; (c) traitement du domaine spatial puis une deuxième branche résiduelle traite le domaine temporel et enfin leurs fusions.

CNN $2D$ appliqués aux images individuelles de la vidéo pour faire la reconnaissance d'action. Enfin, la majorité des modèles $2D$ ont été étendus aux données $3D$, tels que SqueezeNet, MobileNet, MobileNet V2, ShuffleNet et ShuffleNet V2 [71].

Dans le contexte d'analyse des $STIS$, nous trouvons la méthode *DuPLO* [60] qui applique des convolutions $2D$ mais sur une image multi-bandes qui est construite à partir de toutes les images de la série (*i.e.*, une bande par date). Un auto-encodeur convolutif $3D$ est aussi proposé dans [66] pour générer automatiquement des caractéristiques spatio-temporelles à partir d'un $STIS$ à des fins de segmentation.

2.4.3.2 Fusion progressive

Dans notre contexte, les méthodes à fusion progressive sont souvent basées sur l'utilisation de deux réseaux, un sur les données brutes et l'autre sur le flux optique. Puis les caractéristiques extraites sont agrégées et traitées par un autre réseau avant la décision finale. Parmi ces méthodes, il y a celle nommée *Two-Stream Network* [39]. Les auteurs ont étudié deux stratégies de fusion possibles. La première stratégie utilise un seul réseau après la fusion et la deuxième utilise deux couches de fusion, comme représenté sur la figure 2.11. Puis *Two-Stream Network* a été amélioré en proposant un réseau à deux flux [38], l'un est formé sur la vidéo RGB et l'autre sur la vidéo du mouvement (flux optique). Ces réseaux sont reliés via des connexions résiduelles afin d'apprendre l'interaction entre l'apparence et le mouvement. Cette stratégie est adoptée pour les scènes qui sont déformées et qui contiennent des mouvements. Cependant, la limite du *Two-Stream Network* est qu'il traite une seule image de la vidéo. Le modèle *Temporal Segment Networks* [139] outrepasse cette limitation en segmentant la vidéo en plusieurs clips. Puis pour chaque clip, une image

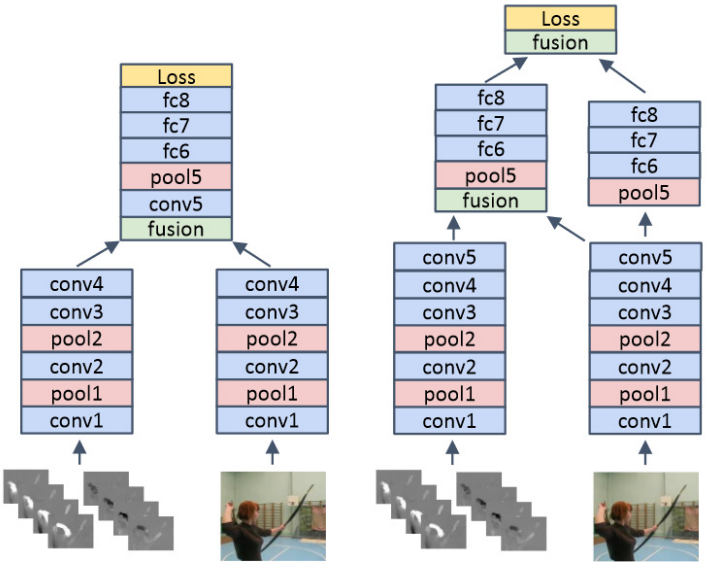


Figure 2.11 – Les deux stratégies de fusion explorées par le *Two-Stream Network* dans [39].

est tirée aléatoirement qui servira d'entrée pour le premier réseau. Le deuxième considère toujours tout le cube du flux optique. Enfin, l'agrégation des caractéristiques extraites est opérée afin de conduire à une décision finale.

Une autre approche, nommée *Representation Flow* [140], s'inspire du flux optique et apprend la représentation du mouvement effectué dans la scène. Cette dernière propose une nouvelle couche qui traite les caractéristiques extraites par un $CNN\ 2D$ de deux images consécutives afin de représenter le mouvement subi entre elles. Une dernière approche prend d'un coté la vidéo RGB brute et de l'autre coté le flux optique. Chaque entrée est traitée avec un $CNN\ 3D$. Cette dernière est nommée *Flow Gated Network* [22]. La fusion des caractéristiques extraites permet de faire le lien entre l'apparence et le mouvement.

2.4.3.3 Fusion tardive

Parmi les méthodes avec une fusion tardive des caractéristiques, nous rencontrons celles qui traitent le temps et le spatial séparément en entraînant deux modèles, un sur chaque domaine, puis les résultats sont agrégés avant la décision finale. Il y a également des méthodes qui prennent en entrée deux types de données, comme la vidéo brute avec la vidéo du flux optique, chaque donnée est traitée via un réseau puis comme précédemment, les caractéristiques sont agrégées.

Pour commencer, nous présentons d'abord quelques méthodes développées dans le cadre d'analyse des données satellitaires pour la cartographie de l'occupation des sols. Certaines méthodes utilisent un réseau mixte avec des convolutions $1D$ et $3D$ qui alternent. Dans ce contexte, le modèle $FG-UNet$ (*fine grained* UNet) est une adaptation du $U-Net$ pour traiter les domaines temporel et spatial des $STIS$ [126]. Notons que les caractéris-

tiques spatio-temporelles profondes sont classiquement apprises de manière supervisée, mais certaines méthodes ont été proposées pour les apprendre de manière non supervisée en utilisant des architectures d'auto-encodeurs [44, 66].

Dans le contexte d'analyse de vidéos, le modèle $VideoGCRF$ [20], basé sur les *Gaussien Conditionnal Random Field* profonds, a été adapté pour la segmentation de vidéo tout en fixant les paramètres du modèle de façon automatique. Ce modèle traite deux images successives de la vidéo puis agrège les informations extraites. Du côté de la classification, la majorité des méthodes se base sur l'utilisation de deux CNN sur deux données différentes. Parmi celles-ci, *Temporal-ConvNet* [81] se compose de deux réseaux de type $ResNet{-}101$. Ces deux réseaux traitent une image de la vidéo et la vidéo du flux optique puis les caractéristiques extraites sont concaténées et sont données soit à un $LSTM$, soit à un $CNN\ 1D$ pour avoir une décision finale. Les auteurs de [97] combinent un $CNN\ 3D$ une fois avec un $CNN\ 1D$ et une fois avec un $CNN\ 2D$. Le $CNN\ 1D$ prend en entrée le signal audio brut. En revanche le $CNN\ 2D$ prend en entrée une image $2D$ du spectrogramme issu du signal audio [93]. Une fusion tardive des caractéristiques déjà extraites est opérée puis elles sont passées à un classificateur de type linéaire pour prédire les émotions des gens dans des vidéos.

2.4.3.4 Nuages de points

La dernière catégorie de méthodes que nous présentons traite des données qui ont une structure de nuages de points [103]. La représentation des données est un ensemble de points dans un espace $3D$. Deux dimensions spatiales plus une troisième qui représente la profondeur. Un point est identifié par ses coordonnées (x, y, z) et d'autres caractéristiques comme par exemple les couleurs [113]. À titre d'exemple les humains peuvent être représentés par les points caractéristiques de leurs squelettes. Les méthodes conçues pour ces données nécessitent une mémoire importante pour traiter une telle quantité de données, en particulier dans le contexte de l'analyse des STI.

Parmi les méthodes construites pour ce type de données, nous trouvons $PointNet$ [103]. Cette dernière effectue un traitement point par point avec un réseau composé de plusieurs couches linéaires. Les points sont considérés dans un ordre aléatoire. Ensuite une agrégation des caractéristiques extraites suivie par un max-pooling sont appliqués afin d'avoir une information globale des données d'entrée. Une amélioration de $PointNet$ a donné naissance à $PointNet++$ [102]. Ce dernier explore l'espace afin d'intégrer une information locale pour chaque point. Pour ce faire, un regroupement des points est d'abord effectué avec la méthode *Farthest Point Sampling*. Ensuite, le modèle $PointNet$ est appliqué sur l'ensemble des points de chaque groupe. En répétant ce processus successivement, des caractéristiques globales sont obtenues à travers les caractéristiques locales. Néanmoins, une telle stratégie requiert beaucoup de ressources de calculs. Les convolutions $3D$ ont été adaptées aux nuages de points en proposant le modèle $PointConv$ [148] qui applique des convolutions sur les arêtes des points. L'apprentissage de ces filtres de convolution se fait avec l'approximation *Monte Carlo*. Ensuite, nous citons le *Dynamic Graph CNN* ($DGCNN$)

[141] où un nouveau module de convolution est proposé, nommé *EdgeConv*. Comme son nom l'indique, la convolution est appliquée sur les poids des arêtes qui sont directement connectées aux nœuds. Pour finir, nous présentons une approche plus récente *Skeleton Points Interaction Learning* ($SPIL$) [128]. Cette méthode traite les points caractéristiques des squelettes humains obtenus par la méthode d'estimation de pose des personnes dans une région ($RMPE$) proposée dans [35]. L'ensemble des points de chaque squelette est considéré comme un graphe. Avec cette stratégie, $SPIL$ peut apprendre l'interaction entre les objets par l'inter-relation des points des graphes.

2.5 Discussion

Ce chapitre a présenté un panorama de différentes méthodes d'extraction de caractéristiques à partir de STI. Ces méthodes sont regroupées en trois catégories principales en fonction de la nature des caractéristiques extraites. Dans chaque catégorie, nous avons présenté des méthodes qui extraient des caractéristiques *hand-crafted* mais également des méthodes qui apprennent les caractéristiques de façon automatique. Les résultats obtenus avec les méthodes purement temporelles sont compétitives sur des données avec un contenu non-déformable comme les $STIS$. Ces résultats peuvent néanmoins être améliorés quand les informations spatiales sont rajoutées, en particulier quand les données traitées ont une forte structure spatiale. Quand les données ont un contenu déformable, comme les vidéos de football ou d'action, ces méthodes deviennent moins pertinentes car si un pixel en un instant t se décale spatialement en $t+1$, alors il n'est pas possible d'analyser l'action réalisée dans la scène. Autrement dit, lors d'un déplacement spatial, il n'y a pas de mise en correspondance entre les pixels à la même position spatiale. L'ajout des informations spatiales de façon « manuelle » ne reste pas naturelle, conduisant à des caractéristiques limitées ou incomplètes.

Les méthodes qui traitent simultanément les domaines spatial et temporel conduisent à des caractéristiques spatio-temporelles. L'extraction de telles informations est une tâche complexe, en particulier avec les méthodes *hand-crafted*. Les réseaux de neurones profonds ont permis d'apprendre automatiquement des caractéristiques spatio-temporelles avec des convolutions $3D$. La difficulté de ces méthodes réside dans l'art de leur conception sans oublier l'entraînement de ces modèles qui peut nécessiter beaucoup de ressources. De plus, l'utilisation de modèles décisionnels contenant des millions de paramètres peut conduire à des solutions difficiles à expliquer et à interpréter, en particulier si l'on considère des données $2D + t$.

Dans les chapitres suivants, nous présentons les contributions proposées durant ce doctorat afin de définir des caractéristiques spatio-temporelles à partir de STI. Ces dernières sont d'un coté *hand-crafted* et de l'autre apprises lors d'une optimisation globale automatique. Le cœur de ces méthodes est basé sur un changement de représentation des données initiales. Les méthodes que nous avons développées ont été appliquées dans deux cadres applicatifs qui ne comportent pas le même type de contenus visuels. Le premier est non-déformable comme les $STIS$ et le second est déformable comme les vidéos. Dans notre cas,

nous avons dans un premier temps étudié la stabilité temporelle en analysant les valeurs répétées successivement en définissant des caractéristiques *hand-crafted*. Dans un deuxième temps, nous avons enrichi les pixels temporels avec une information spatiale en nous basant sur des courbes dans le domaine spatial qui sont générées de différentes manières. Cela a donné naissance à une représentation planaire qui permet a un $CNN\ 2D$ d'apprendre des caractéristiques spatio-temporelles.

Chapitre 3

Cadre applicatif : télédétection et vidéo

Un problème sans solution est un problème mal posé.

– Albert Einstein

Dans ce chapitre, nous décrivons les deux cadres applicatifs qui sont abordés dans cette thèse et nous présentons les jeux de données utilisés dans les expérimentations. Deux types de données sont considérés. Le premier type est relatif aux Séries Temporelles d'Images Satellitaires ($STIS$). Le deuxième type est lié à des vidéos issues de caméras de sécurité ou de films.

3.1 Introduction

Afin d'évaluer la généricité et l'intérêt des méthodes proposées, nous considérons deux cadres applicatifs différents. Le premier est une application de télédétection qui consiste à

analyser la couverture des sols à partir de $STIS$ à des fins de cartographie automatique. La deuxième application concerne un problème de classification de vidéos pour la détection de scènes de violence. Il est à noter que les données de ces deux applications ne sont pas de même nature. Les données de télédétection sont des images avec un contenu non-déformable qui évolue dans le temps, c'est-à-dire que le capteur acquiert la même scène à des instants différents. Pour les vidéos, ces dernières ont un contenu déformable qui peut être dû au déplacement des objets dans la scène ou au mouvement de la caméra.

Les méthodes développées dans cette thèse prennent en entrée une séquence temporelle d'images (STI), notée $S_{images}=< I_1, \ldots, I_T >$ constituée de T images. Les images de la STI sont toutes définies dans le même domaine spatial $\mathcal{D} = [\![1, \mathbb{W}]\!] \times [\![1, \mathbb{H}]\!]$ où $\mathbb{W}$ et $\mathbb{H}$ représentent respectivement leur largeur et hauteur.

Un pixel temporel $(p_t)_{t=1}^T$ est l'ensemble des pixels qui sont à la même position spatiale à travers le temps T. Il est représenté sous la forme d'une fonction définie comme :

$$\begin{array}{rccc} (p_t)_{t=1}^T : & \mathcal{D} & \rightarrow & (\mathbb{R}^B)^T \\ & (x,y) & \mapsto & (I_t(x,y,1), \ldots I_t(x,y,B))_{t=1}^T \end{array} \tag{3.1}$$

qui associe à un pixel de coordonnées (x, y) une série temporelle de valeurs colorimétriques, où chaque valeur peut prendre la forme d'un scalaire ou d'un vecteur qui dépend du nombre de bandes spectrales B des images.

Les deux applications sont ensuite introduites dans les sections 3.2 et 3.3. Pour chacune des applications, les données et les problématiques sont évoquées.

3.2 Analyse de séries temporelles d'images satellitaires

Dans le cadre de ces travaux, nous nous sommes focalisés dans un premier temps sur l'analyse de $STIS$ pour la télédétection. Pour ce faire, le choix des données à traiter reste important en fonction de l'application choisie. Par exemple, l'analyse du cycle de végétation nécessite une $STIS$ avec une grande fréquence temporelle. Plusieurs satellites acquièrent des $STIS$ mais chacun a ses propres caractéristiques. Par exemple SPOT-4 a été lancé en 1998 mais capte une image tous les 26 jours avec une résolution spatiale de 20 mètres. En 2013, un autre satellite nommé Landsat-8 a permis d'acquérir davantage d'images (chaque 16 jours) mais avec une résolution spatiale de 30 mètres. Entre 2015 et 2017, la *European Space Agency* (ESA) a lancé deux satellites de télédétection dans le contexte du programme européen de surveillance de la Terre **Copernicus** qui est l'ex-programme *Global Monitoring for Environment and Security* (GMES) [7, 33]. Les deux satellites sont nommés Sentinel-2A et Sentinel-2B. Ils ont pour mission de fournir des $STIS$ avec une revisite temporelle de cinq jours, une résolution spectrale de 13 bandes dans le visible et l'infra-rouge et une résolution spatiale allant de 10 à 60 mètres suivant la longueur d'onde.

3.2.1 Applications thématiques de télédétection

Deux applications thématiques sont ciblées dans cette étude. La première concerne un problème d'analyse de la couverture urbaine et la deuxième se focalise sur l'analyse de différentes parcelles agricoles pour contrôler les pratiques de gestion des cultures. Dans la suite, nous détaillons les deux applications. Nous présentons ensuite les données utilisées lors de nos expérimentations et leurs vérités terrain associées pour l'apprentissage et l'évaluation des résultats obtenus.

3.2.1.1 Analyse de la couverture urbaine

La couverture urbaine modélise l'emprise du tissu urbain bâti. Autrement dit, cette couverture est relative aux zones artificielles ou occupées par des constructions humaines telles que les bâtiments, les maisons individuelles ou les infrastructures de transports, etc. Dans cette application, nous souhaitons automatiser la détection de l'étendue de la couverture urbaine. Cela permet par exemple d'analyser la croissance de la population en fonction de la croissance de l'étalement urbain. Cela peut également permettre de faciliter des études d'urbanisme pour comprendre des dynamiques d'évolution des villes.

3.2.1.2 Analyse des parcelles agricoles

L'analyse des parcelles consiste à examiner la couverture des sols cultivés afin d'aider les décideurs politiques en matière d'agriculture et d'environnement. Cela peut aider par exemple à contrôler les pratiques de gestion agricoles à grande échelle pour vérifier les déclarations annuelles des agriculteurs. Les données utilisées sont extraites d'une base géographique servant de référence à l'instruction des aides de la politique agricole commune (PAC). Cette base est le registre parcellaire graphique (RPG [1]). Elle contient les données graphiques des parcelles avec leur culture principale. Dans notre cas, nous nous limitons à quatre classes thématiques de parcelles qui sont :

- les prairies ;
- les vignes ;
- les vergers traditionnels ;
- les vergers intensifs.

Un tel choix du jeu de données est lié au contexte d'un projet de recherche pluridisciplinaire national français ANR TIMES [2] où les données ont été sélectionnées par un consortium de géographes. Ces quatre classes thématiques sont très complexes à identifier car certaines parcelles sont soumises à plusieurs pratiques de gestions agricoles en fonction des saisons

1. http://professionnels.ign.fr/rpg
2. TIMES project – *High-performance processing techniques for mapping and monitoring environmental changes from massive, heterogeneous and high frequency data times series*, lien https://anr.fr/Projet-ANR-17-CE23-0015

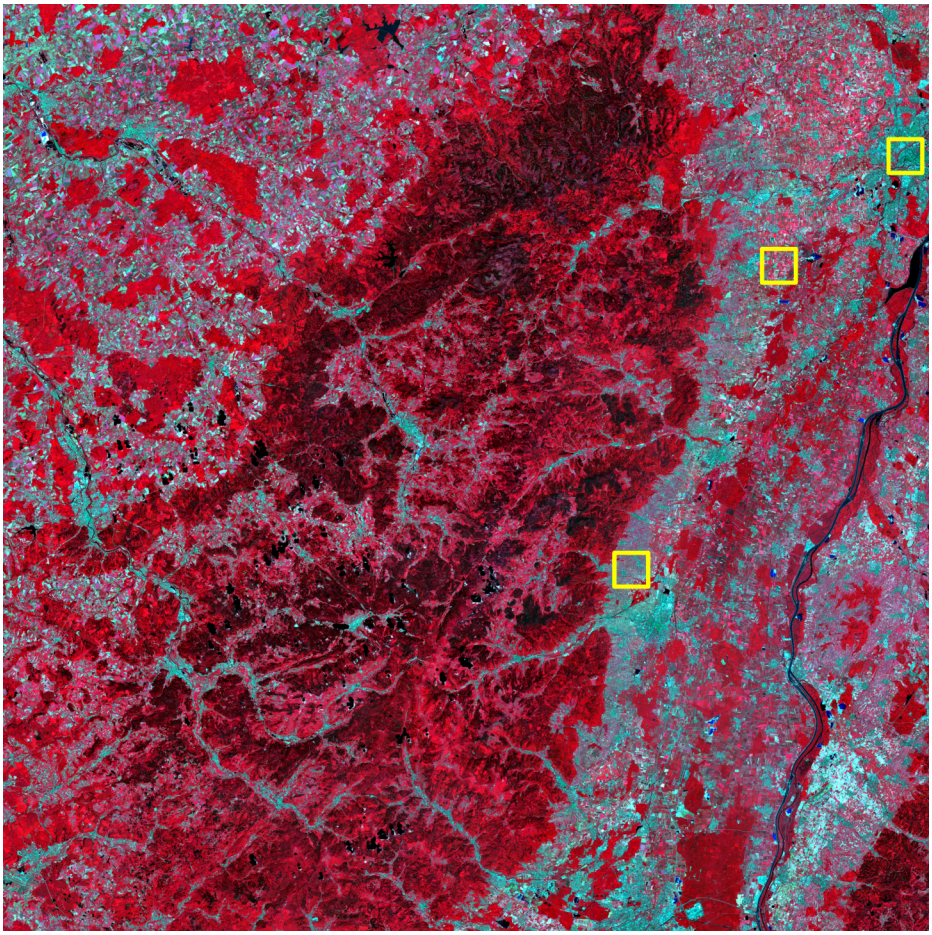


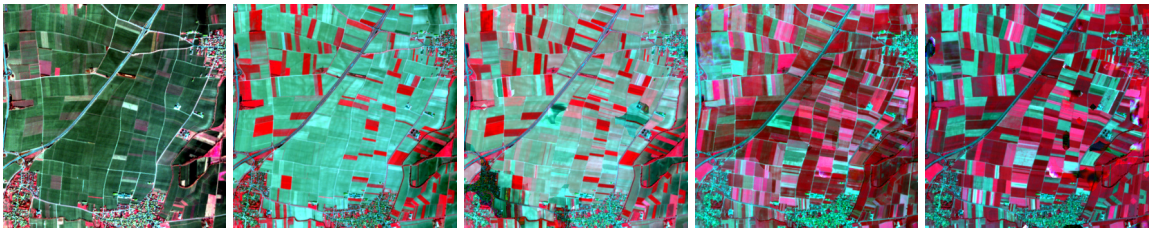
Figure 3.1 – Illustration de la tuile 32ULU. Les trois zones encadrées correspondent respectivement à trois zooms dans différentes zones qui sont présentées dans la figure 3.2 (trié du haut vers le bas).

et des politiques de gestion du territoire. Par exemple, les arbres sont plus alignés et ce de façon régulière dans les vergers intensifs alors qu'ils peuvent être éparpillés aléatoirement sur toute la parcelle dans les vergers traditionnels. Des études précédentes ont également mis en évidence que les vergers sont des classes ambiguës et ont tendance à être confondus avec les prairies dans la plupart des méthodes de classification de l'état de l'art[126, 32]. Pour différencier ces classes, il serait important de prendre en considération le domaine spatial en analysant les $STIS$.

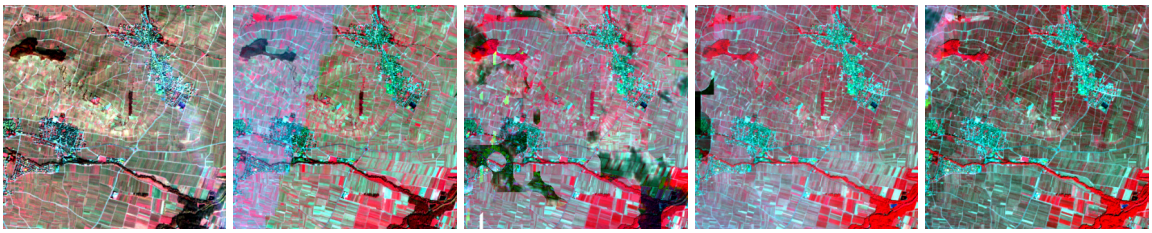
3.2.2 Données et vérité terrain

La séquence temporelle d'images utilisée dans la première application est une $STIS$ composée de 50 images qui sont acquises par Sentinel-2 durant l'année 2017. Les images de la série couvrent la même zone géographique dans l'Est de la France, précisément dans la tuile 32ULU. Généralement, les images satellitaires sont confrontées à des perturbations

(a) Zoom sur une zone urbaine de la ville de Strasbourg.



(b) Zoom sur une zone agricole présentant des prairies.



(c) Zoom sur une zone agricole présentant des vignes.

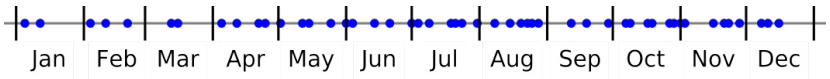


Figure 3.2 – Illustration de quelques images de la série dans différentes zones avec la distribution des 50 images de la $STIS$ couvrant l'année 2017. Chaque ligne correspond à une des trois zones encadrées dans la figure 3.1 (trié du haut vers le bas).

au moment de l'acquisition. Un pré-traitement est alors appliqué sur chaque image afin de réduire les problèmes perturbateurs. Les premières corrections sont géométriques et radiométriques afin de rendre respectivement les images de la série superposables et aussi radiométriquement comparables, c'est-à-dire qu'un pixel couvre la même zone géographique le long de la série et que les valeurs de reflectance sont comparables d'une image à l'autre. Ce processus a été appliqué grâce à la collaboration du CNES, du CESBIO et du DLR qui a conduit au développement de la chaîne MAJA [3][48]. La chaîne MAJA fournit également les masques de nuages, d'ombres et de saturation associés à chaque image de la série.

La dernière étape de pré-traitement consiste à reconstruire les données manquantes à cause de la présence des nuages et aussi des problèmes de saturation. La solution que nous avons utilisée est l'application d'une interpolation linéaire sur les pixels temporels. Le principe de cette méthode consiste à remplacer la valeur manquante du pixel, par une valeur calculée en faisant l'hypothèse que l'évolution entre deux valeurs correctes consécutives

3. https://labo.obs-mip.fr/multitemp/maccs-comment-ca-marche/

de pixels est monotone. De cette façon, nous garantissons la même longueur pour toutes les séries temporelles $(p_t)_{t=1}^{T}$ de la $STIS$. La figure 3.1 présente l'image de toute la tuile 32ULU avec trois zones sélectionnées dans différentes régions (encadrées avec un rectangle jaune). La figure 3.2 présente un zoom sur les trois zones sélectionnées avec quelques images ponctuelles de la $STIS$. La distribution temporelle des images durant l'année 2017 est aussi affichée dans la figure 3.2. Par la suite, nous avons sélectionné seulement les bandes qui sont à une résolution spatiale de 10 mètres, ce qui conduit à n'utiliser que les bandes spectrales qui sont le proche infra-rouge Nir, le rouge R, le vert G et le bleu B.

Deux données de vérité terrain sont utilisées, une pour chacune des deux applications de l'analyse de la couverture urbaine et l'analyse des parcelles.

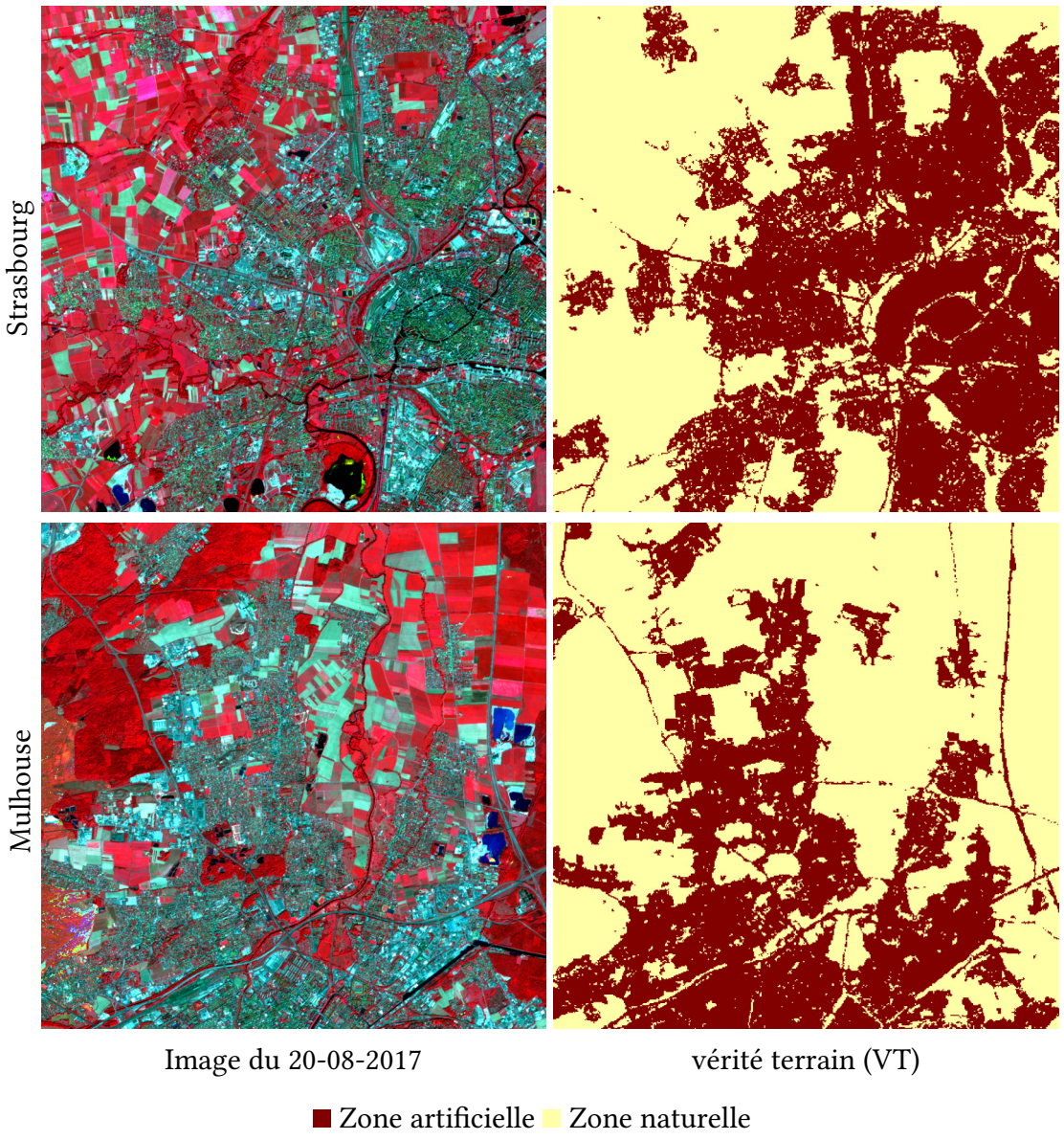


Figure 3.3 – Illustration des images de deux zones avec leurs vérités terrains (VT).

3.2.2.1 Analyse de la couverture urbaine

Deux zones géographiques urbaines différentes sont sélectionnées. La première zone se focalise sur la ville de Strasbourg et la deuxième couvre celle de Mulhouse. La première colonne de la figure 3.3 montre ces deux zones géographiques. Chaque image a une dimension de 1000×1000 pixels. En plus des deux $STIS$, nous disposons d'un produit d'imperméabilité. Ce dernier représente le pourcentage d'imperméabilisation du sol. Le produit d'imperméabilité est le résultat d'un projet du programme européen Copernicus publié par l'Agence européenne pour l'environnement [4]. Il définit l'imperméabilité des matériaux et est fourni au niveau du pixel. La résolution spatiale de ce produit est de 20 mètres, mais nous l'avons ré-échantillonné à 10 mètres pour l'adapter à la résolution spatiale des images Sentinel-2. Chaque valeur de pixel dans ces données de référence estime un degré d'imperméabilité (0–100%). Dans cette étude thématique, nous avons utilisé ce produit comme vérité terrain (VT) pour distinguer les zones naturelles (imperméabilité 0%) des zones artificielles (imperméabilité $> 0\%$). La deuxième colonne de la figure 3.3 illustre les données de référence d'imperméabilité pour les deux zones géographiques.

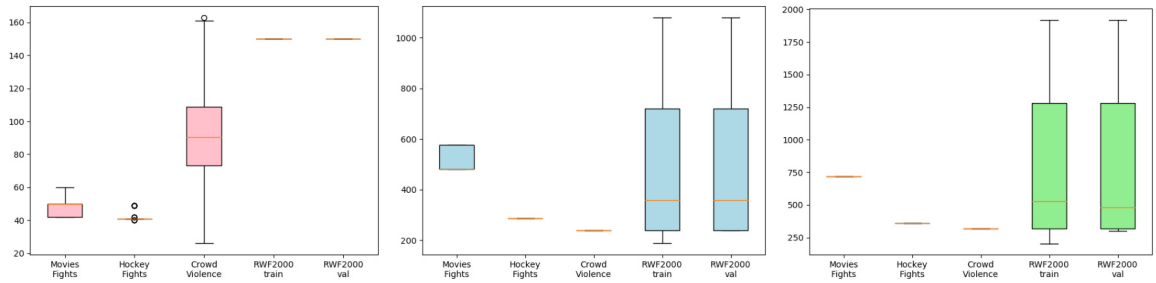
3.2.2.2 Analyse des parcelles agricoles

Les données de référence utilisées dans cette application sont extraites du RPG. Une photo-interprétation est appliquée pour corriger, si nécessaire, les délimitations des parcelles. Les données du RPG sont enregistrées sous la forme vectorielle. Par la suite, nous avons rasterisé le RPG à la même résolution spatiale que les images Sentinel-2 (10 m) et nous n'avons conservé que les polygones correspondant aux quatre classes étudiées. Nous avons également ignoré les parcelles d'une taille inférieure à 12 pixels. Le tableau 3.1 présente le nombre de polygones conservés pour notre étude.

Tableau 3.1 – Nombre de parcelles agricoles collectées dans chaque classe avec des informations statistiques associées.

Classes	**# poly.**	**aire (en pixels)**	
		moyenne	écart-type
Prairies	1 045	250	338
Vignes	562	50	47
Vergers traditionnels	136	154	305
Vergers intensifs	191	129	115
Total	1 934	–	–

4. https://land.copernicus.eu/



(a) Nombre d'images (b) Hauteur (c) Largeur

Figure 3.4 – Distribution statistique des dimensions spatiales et de la durée des vidéos.

3.3 Analyse de vidéos

En analyse de vidéos, notre but principal est de réaliser un système qui classifie le contenu d'une vidéo de façon automatique. Cela peut être réalisé dans différentes applications d'analyse de vidéos comme la reconnaissance d'actions, d'émotions ou d'objets. Dans notre cas, nous nous intéressons à un problème particulier de reconnaissance d'actions, plus précisément, la reconnaissance de scènes de violence. Cette tâche est considérée comme un problème classique de classification globale de vidéos. La reconnaissance de la violence est étudiée pour pouvoir sécuriser les espaces publics sous surveillance, par exemple les gares ou les prisons.

3.3.1 Bases de vidéos

Dans le contexte de la détection de la violence, de nombreux chercheurs ont proposé différents jeux de données de vidéos. Les trois bases les plus connues sont *Movies Fight* [91], *Hockey Fight* [91] et *Crowd Violence* [49]. Récemment, la base *RWF2000* [22] a été proposée. Elle contient plus de données que les bases précédentes. Nous allons utiliser dans notre étude expérimentale ces quatre jeux de données représentatifs qui sont décrits ci-dessous :

- ***Movies Fight*** [91] est l'une des premières bases de vidéos conçue pour l'analyse de violence. Elle est constituée de 200 vidéos collectées à partir de différentes scènes de films. Elle contient un nombre de vidéos équilibré pour chacune des deux classes : 100 vidéos violentes et 100 vidéos non violentes ;
- ***Hockey Fight*** est une deuxième base proposée par les auteurs de *Movies Fight* [91]. Les vidéos de cette dernière sont issues de matchs de *Hockey* d'une ligue nationale. Le nombre de vidéos collectées est largement supérieur à celui de la précédente base. Il est égal à 1000. Chaque classe comprend 500 vidéos ;
- ***Crowd Violence*** [49] est une base plus complexe que les deux précédentes car les vidéos sont toutes issues de foules en mouvement. La plupart d'entre elles sont extraites de matchs de football. La seule limite de cette base est le faible nombre de

vidéos récoltées qui est égal à 246. Le nombre de vidéos par classe est d'environ 123 ;

- ***RWF2000*** [22] est une base plus récente. Les vidéos de cette dernière sont toutes collectées à partir de YouTube avec des requêtes par mots-clés dans différentes langues. Les auteurs ont ainsi collecté 2000 vidéos capturées à partir d'une caméra de surveillance. Chaque classe contient 1000 vidéos. La base est divisée par les auteurs en deux ensembles où l'un représente l'ensemble d'entraînement alors que le deuxième est relatif à celui du test.

Les bases présentées n'ont pas les mêmes caractéristiques en termes de dimensions et de durées. La figure 3.4 présente des boîtes à moustaches qui donnent une information sur la hauteur, la largeur et le nombre d'images des vidéos de chaque base. *Crowd Violence* [49] est la base qui a les vidéos les plus petites (H et W les plus basses) mais avec des durées très variables. Les vidéos de *Hockey Fight* [91] partagent aussi la même taille avec une durée temporelle qui est presque égale. Les bases *Movies Fight* [91] et *RWF2000* [22] ont des vidéos qui ne partagent pas la même taille mais les vidéos de *RWF2000* ont toutes la même durée.

3.3.2 Difficultés

Nous rappelons que la tâche liée à cette application est la classification binaire de vidéos (violente ou non violente). Il est cependant difficile d'arriver à avoir un bon résultat car la scène observée peut s'avérer très complexe. La figure 3.5 montre quelques exemples de chacune des bases de vidéos. La base *Movies Fights* est celle qui présente les vidéos les plus simples comme nous pouvons le voir sur la figure 3.5. Toutefois ce n'est pas le cas pour l'autre base proposée par les mêmes auteurs *Hockey Fights*. Dans les deux exemples de cas « violents », nous pouvons aisément dire que la vidéo de gauche est violente mais cela est plus subjectif pour celle qui est à droite. La base *Crowd Violence* est la plus complexe à cause du mouvement de foule qui reste très difficile à analyser. *RWF2000* propose une quantité de vidéos plus élevée qui présentent un point commun, toutes les vidéos proviennent de caméras de surveillance. La différence entre ces sources de données et les caméras utilisées par les humains repose sur le mode d'acquisition. Les caméras de surveillance sont fixes avec seulement des scènes qui contiennent du mouvement. Les vidéos de *RWF2000* sont prises dans les rues, dans les restaurants et même dans les ascenseurs. L'analyse dans cette base peut être complexe dans ce type de cas comme par exemple une personne agressée dans un ascenseur où la victime peut être cachée par une autre personne (phénomène d'occlusion).

3.4 Discussion

Les deux cadres applicatifs présentés dans ce chapitre sont relatifs à l'analyse de données $2D + t$. La première application concerne les séries temporelles d'images satellitaires dans le but d'analyser la couverture urbaine et la classification de parcelles agricoles. Les objets d'intérêt y sont relativement peu déformables. La deuxième application cible la classification des vidéos pour la reconnaissance de la violence. Les objets d'intérêt y sont rela-

tivement déformables dans le temps. Le but principal de la thèse est l'extraction de caractéristiques spatio-temporelles qui peuvent être *hand-crafted* ou apprises automatiquement. Ces caractéristiques sont ainsi impliquées pour résoudre les problématiques de chacune des deux applications.

Movies Fights

(a) Violente (b) Non violente

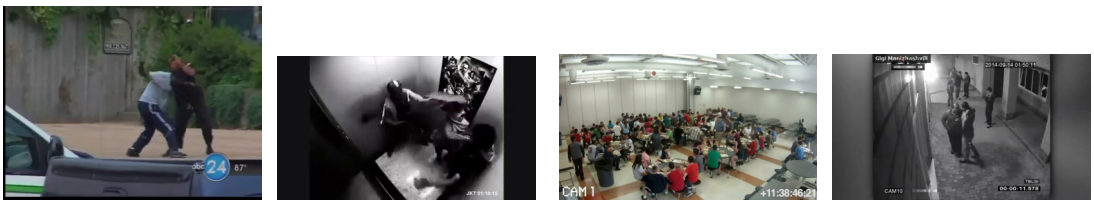
Hockey Fights

(c) Violente (d) Non violente

Crowd Violence

(e) Violente (f) Non violente

RWF2000



(g) Violente (h) Non violente

Figure 3.5 – Exemples d'images extraites de vidéos violentes et non violentes de chacune des bases dans le cadre de l'application d'analyse de vidéos.

Chapitre 4

Étude de la stabilité

Le doute est le commencement de la sagesse.

– Aristote

Dans ce chapitre, nous nous intéressons à l'étude du complémentaire du changement qui est la stabilité dans le contexte de l'analyse d'une STI. La section 4.1 évoque quelques méthodes d'analyse de changements. La section 4.2 présente la mesure de stabilité proposée et le changement de représentation considéré pour la calculer. Comme la stabilité repose sur l'égalité des valeurs successives, la section 4.3 détaille un descriptif de la notion d'égalité avec différentes stratégies amenant à définir une stabilité spatio-temporelle. La section 4.5 propose une composition des caractéristiques extraites pour générer une image en fausse couleur qui résume la STI. Les résultats qualitatifs et quantitatifs de l'étude expérimentale sont exposés dans la section 4.6. Enfin, le bilan scientifique est présenté dans la section 4.7.

4.1 Introduction

En vision par ordinateur, l'analyse de changements est une problématique majeure. Elle permet à partir de la même scène acquise à des instants différents, la détection des régions qui n'ont pas le même aspect dans plusieurs images. Plusieurs méthodes se basent sur l'analyse de changements afin de trouver des régions ou des objets d'intérêt. Par exemple, la localisation et le suivi d'objets peuvent concerner un piéton ou une voiture traversant la scène en télé-surveillance [120], l'analyse des dommages provoqués par les catastrophes naturelles, l'identification de l'évolution des chantiers en télédétection [25] ou encore le suivi d'une pathologie en médecine. La figure 4.1 illustre un exemple de détection de changements issu de [120]. Les auteurs de [120] ont pour objectif de détecter le mouvement des gens en premier plan dans un endroit télé-surveillé.

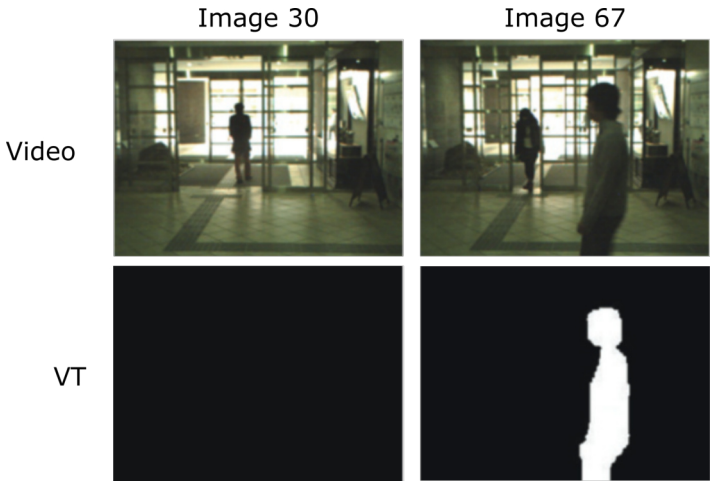


Figure 4.1 – Exemple du changement subi en premier plan dans deux images de la même vidéo avec leurs vérités terrain (VT) associées [120]. Les deux VT présentent uniquement le changement du premier plan.

Historiquement, les méthodes d'analyse de changements consistaient à procurer des indices permettant la compréhension de la scène acquise. La majorité des méthodes ne traitent que le domaine temporel en considérant que chaque pixel est un ensemble de mesures ordonnées chronologiquement [62, 15, 99, 109]. Toutes ces méthodes n'incluent pas d'informations spatiales au cours de l'analyse. En analysant le changement des pixels temporels, il est possible de les étiqueter selon leurs évolutions temporelles. Par exemple en télédétection, le cycle d'évolution d'un pixel temporel permet de savoir son type de culture. Dans le même contexte, les pixels localisés dans les zones urbaines restent stables dans le temps. Dans le cas des vidéos, l'arrière plan dans les émissions de télévision reste lui aussi stable si la caméra fixe la même scène. Dans ce contexte, il est intéressant d'aborder l'analyse de changement de manière duale en analysant la stabilité. En recherchant le nombre de périodes stables à travers le temps, il est possible de concevoir une solution pour différents problèmes comme l'analyse de la tache urbaine dans le cadre applicatif de télédétection.

Par la suite, ce chapitre se focalise sur l'analyse de la stabilité spatio-temporelle d'une STI conduisant à une nouvelle représentation des données. Celle-ci est utilisée pour extraire des caractéristiques spatio-temporelles de stabilité et les utiliser par la suite pour résoudre deux problèmes différents qui sont l'analyse de la tache urbaine et la détection de violence à partir de vidéos. En analyse de vidéo, certaines méthodes consistent à résumer les vidéos en gardant seulement les images les plus représentatives [96]. Contrairement à ces méthodes, nous nous proposons de résumer la STI en une seule image en fausses couleurs contenant différentes informations complémentaires de stabilité. Un tel résumé permet de visualiser une seule image au lieu de toutes les images de la STI.

4.2 Mesure de la stabilité

Une mesure de la stabilité est un moyen d'étudier l'homogénéité de l'évolution d'une scène à travers le temps. Cette homogénéité tient compte des différentes modifications temporelles qui peuvent avoir lieu. La figure 4.2 présente deux exemples de pixels temporels extraits de deux zones différentes d'une $STIS$. Les valeurs de ces séquences sont caractérisées par leurs indices de végétation $NDVI$. Le pixel du haut est localisé dans une zone agricole, quant au deuxième, il est localisé dans une zone urbaine. L'évolution temporelle de ces deux pixels n'est pas la même car la végétation pousse au cours du temps. Nous observons plus de stabilité temporelle pour le pixel situé dans la zone urbaine. Cet exemple présente l'intérêt de se focaliser sur la stabilité au lieu du changement.

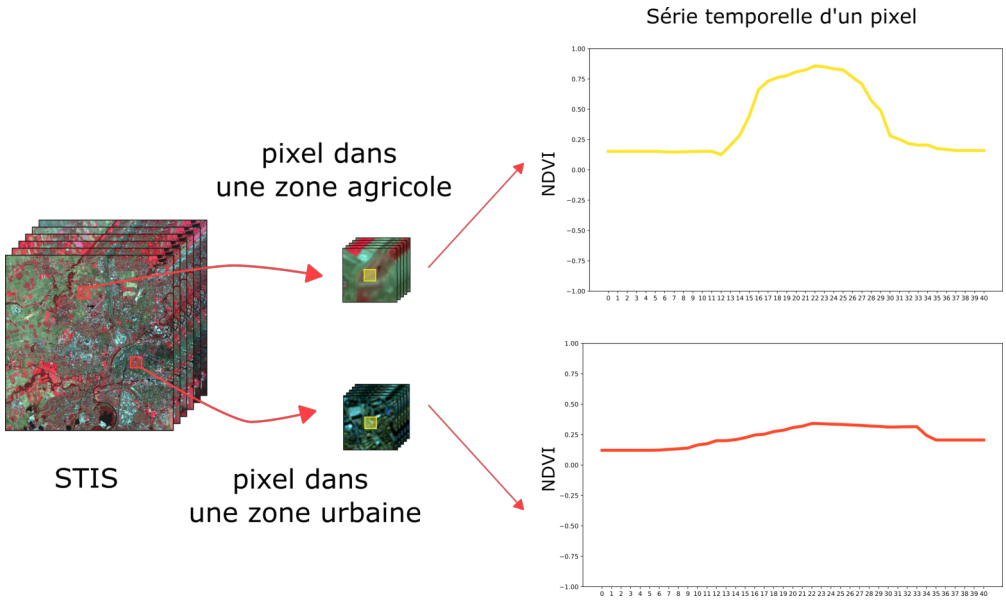


Figure 4.2 – Illustration de deux pixels temporels dans deux zones différentes extraites d'une $STIS$ de Sentinel-2. Le premier est situé dans une zone agricole et le deuxième dans une zone urbaine.

Avant de passer à l'extraction des caractéristiques de stabilité, nous présentons une étape supplémentaire qui va transformer les données originales de la STI vers une nouvelle représentation. Les caractéristiques sont ensuite extraites par l'intermédiaire de cette représentation. Dans la suite, nous présentons la méthode de transformation de la STI vers une nouvelle représentation et les différentes caractéristiques *hand-crafted* extraites.

4.2.1 Nouvelle représentation

La stabilité désigne le maintien d'un état (*e.g.*, valeur colorimétrique du pixel) pendant une certaine plage (temporelle ou spatiale). Pour pouvoir extraire des caractéristiques de stabilité, nous changerons la représentation de la STI vers une nouvelle représentation intermédiaire qui repose sur la mesure de la stabilité dans le temps. Pour ce faire, il faut tout d'abord définir une façon d'étudier la répétition des valeurs successives dans le temps.

En théorie de l'information, l'analyse de la redondance des données est une problématique importante [132]. Le but principal est de compresser les données tout en ayant un contenu le plus identique possible au contenu initial lors de la décompression. Différents algorithmes de compression existent et sont divisés en deux groupes. Les algorithmes du premier groupe compressent les données sans perdre aucun détail. Par contre, les algorithmes du deuxième groupe perdent de l'information au moment de la compression mais le contenu décompressé est presque identique à l'original. Parmi les méthodes de compression sans perte, nous citons le codage de Huffman [57] qui représente les données avec des codes à longueurs variables. Les longueurs des codes sont déterminées à partir des probabilités d'apparition du contenu original des données. Un autre algorithme de compression est LZW (pour *Lempel-Ziv-Welch*) [144]. Ce dernier est basé sur l'utilisation de dictionnaire qui permet de remplacer des chaînes par des symboles. Enfin, il existe l'algorithme *Run Length Encoding* (RLE) [43] qui explore la redondance des valeurs successives en notant un couple (valeur/fréquence absolue).

En morphologie mathématique, certaines méthodes étudient la relation de connectivité des pixels pour définir des zones quasi-plates [124]. Ces dernières sont basées sur l'égalité des valeurs de pixels dans le domaine spatial. Pour ce faire, les pixels d'une image sont considérés comme un graphe non orienté où les arrêtes définissent la relation entre les pixels liés. Cette relation dépend de la différence entre les valeurs de deux pixels liés. En appliquant un critère d'égalité sur ces relations, une segmentation peut être obtenue. D'autres travaux ont étudié les zones quasi-plates sur des vidéos [143] amenant à produire des zones quasi-plates spatio-temporelles.

Dans notre cas, nous avons choisi l'algorithme RLE [43]. Le codage de Huffman et LZW ne peuvent pas être utilisés car ils n'explorent pas la notion de redondance successive. L'algorithme RLE a déjà été utilisé pour l'analyse de séries temporelles dans [106, 3]. Le RLE est un algorithme de compression sans perte qui compresse un pixel temporel $p = (p_t)_{t=1}^{T}$ de longueur T en enregistrant à la fois la valeur et le nombre de répétitions successives de cette valeur. Dans notre cas, nous ne considérons que la deuxième information

et nous omettons la valeur dans le but de ne conserver que l'information de redondance qui représente la stabilité du pixel. Le vecteur résultant est ici noté $RLE(p)$ et sa longueur est notée l_p. La figure 4.3 illustre le calcul de RLE sur un pixel temporel p.

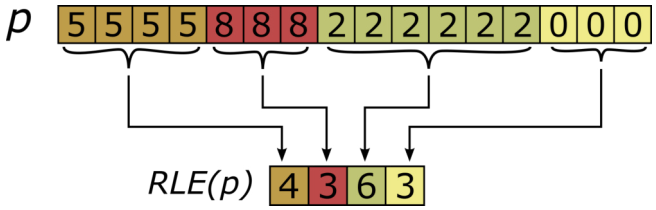


Figure 4.3 – Transformation d'un pixel temporel p basée sur le résultat de l'algorithme *Run Length Encoding* (RLE) [43].

Notre stratégie consiste ici à utiliser l'algorithme RLE pour transformer une STI en une structure moins volumineuse conduisant à une représentation des données dans laquelle différentes informations sur la stabilité peuvent être plus facilement mesurées. Tout d'abord, l'application de l'algorithme RLE se fera au niveau de tous les pixels temporels p. Maintenant que la méthode de transformation est choisie, nous passons à l'extraction de caractéristiques proprement dites.

4.2.2 Définition de caractéristiques

Grâce à la compression de l'algorithme RLE, la STI est transformée vers une nouvelle représentation moins volumineuse grâce à laquelle nous avons défini trois caractéristiques. Ces dernières sont la **stabilité maximale**, le **début de la stabilité maximale** et le **nombre de changements**. Chacune de ces caractéristiques est détaillée par la suite.

4.2.2.1 Stabilité maximale

Étant donné un pixel temporel p, cette première caractéristique capte la plus longue durée où l'attribut du pixel reste stable (*i.e.*, sans changement) dans la série temporelle, notée MS comme *Max Stability*. Elle peut être interprétée de deux façons différentes. Dans le cas des vidéos, elle est exprimée en terme de nombre d'images, sinon dans d'autres cas comme l'analyse des $STIS$, elle est exprimée en terme de différence de dates. La valeur d'une telle caractéristique peut être calculée à l'aide de la fonction définie dans l'équation 4.1 représentée ci-dessous :

$$MS(p) = \|RLE(p)\|_\infty \tag{4.1}$$

où $\|.\|_\infty$ représente la norme infinie.

4.2.2.2 Début de la stabilité maximale

La deuxième caractéristique, notée MSS pour *Max Stability Start*, est directement liée à la stabilité maximale (MS). Elle correspond au début de la période de stabilité maximale d'un pixel p à travers le temps. Cette caractéristique fournit une information qui permet la discrimination des différentes évolutions des pixels. Par exemple, si un pixel reste stable le long de la STI alors la valeur MSS est faible car la MS commence très tôt. Un tel exemple peut être trouvé dans une région urbaine lors de l'analyse des $STIS$ ou dans une zone d'arrière plan lors de l'analyse de vidéos de journal télévisé. La valeur de cette caractéristique peut être calculée de la manière suivante :

$$MSS(p) = \left(\sum_{i=1}^{t_0-1} RLE(p)_i \right) \quad \text{avec} \quad t_0 = \underset{t \in [1, l_p]}{min} \{ t / RLE(p)_t = MS(p) \} \tag{4.2}$$

où t_0 est l'indice de commencement de $MS(p)$.

4.2.2.3 Nombre de changements

La troisième caractéristique, notée NB, correspond au nombre de plages de stabilité trouvées par le RLE. En d'autres termes, elle représente le nombre de changements du pixel temporel $p = (p_t)_{t=1}^{T}$ étudié de longueur T. Soit $RLE(p)$ le vecteur résultant de longueur l_p. La définition de la caractéristique NB se fait via la formule 4.3 présentée ci-dessous :

$$NB(p) = l_p - 1 \tag{4.3}$$

L'algorithme $RLE(p)$ est basé sur l'égalité entre les valeurs successives. Cependant, les images composant la STI ne sont pas acquises au même temps t. Malgré la normalisation, la variabilité des valeurs d'intensité des pixels peut être importante le long de la série puisque les distributions des valeurs des pixels ne sont pas toujours dans la même dynamique d'une image à l'autre. Cela peut être dû aux différentes conditions telles que l'éclairage ou le mouvement. Pour traiter cette question, la notion d'égalité doit être soigneusement étudiée afin d'évaluer, de manière plus réaliste, si deux valeurs de pixels, successives dans le temps, peuvent être considérées comme égales ou non. Nous nous intéressons ainsi par la suite à cette notion d'égalité.

4.3 Notion d'égalité

L'égalité est une relation d'équivalence binaire qui compare deux objets d'un même ensemble E. Ils sont considérés comme identiques si un prédicat donné P est vrai. Dans notre cas, le prédicat P est défini comme une fonction qui teste l'égalité définie sur E, à valeur booléenne définie dans $\mathbb{B}$ (le domaine de définition des valeurs booléennes). La

formulation mathématique de P est présentée dans la fonction 4.4 donnée ci-dessous :

$$\begin{array}{rcccl} P & : & E \times E & \rightarrow & \mathbb{B} \\ & & o_1,\ o_2 & \rightarrow & \begin{cases} Vrai & \text{si} \quad o_1 = o_2 \\ Faux & \text{sinon} \end{cases} \end{array} \tag{4.4}$$

Lorsqu'une STI est considérée, les valeurs des pixels considérés peuvent être continues, discrètes dans un intervalle tel que $[0, 255]$ ou vectorielles lorsqu'un hypercube est pris en compte (les images hyperspectrales [5]). Dans ce contexte, l'égalité des valeurs n'est pas toujours significative car il est incertain d'avoir une répétition successive d'une même couleur. Pour résoudre ce problème, nous proposons de quantifier les valeurs des pixels. Afin d'éviter tout biais sur les nouvelles valeurs représentant la STI après la quantification, elle ne doit pas être appliquée à chaque image I_t, ni au pixel temporel p, mais elle doit être faite au niveau global de tous les pixels qui constituent la STI. De cette façon, la quantification va être adaptée à la distribution des valeurs de la STI et à la nature des caractéristiques utilisées.

Pour effectuer la quantification, nous proposons d'appliquer un algorithme de regroupement des valeurs qui permet de définir des centroïdes optimaux. Nous avons choisi l'algorithme k-Moyenne pour quantifier les valeurs des pixels en seulement k_{quanti} valeurs. k_{quanti} est un paramètre de la méthode qui désigne le nombre de groupes. Ensuite, les valeurs des pixels sont remplacées par les valeurs des centroïdes des classes auxquelles ils appartiennent. Ainsi, une nouvelle STI quantifiée $(J_t)_{t\in[\![1,T]\!]}$ est définie. Nous avons choisi l'erreur quadratique moyenne (EQM) pour évaluer la qualité des données après la quantification. Cette dernière se calcule entre la STI originale $(I_t)_{t\in[\![1,T]\!]}$ et la STI quantifiée $(J_t)_{t\in[\![1,T]\!]}$ [29], définie comme suit :

$$EQM(I, J) = \frac{1}{T \times W \times H \times B} \sum_{t=1}^{T} \sum_{x=1}^{H} \sum_{y=1}^{W} \sum_{b=1}^{B} (I_t\,(x, y, b) - J_t\,(x, y, b))^2 \tag{4.5}$$

Nous avons comparé la méthode de quantification proposée avec une autre méthode naïve. Celle-ci consiste à diviser l'espace des valeurs en k plages de longueurs fixes (inadaptation du contenu des données contrairement au cas du k-Moyenne). Puis, à chaque pixel, nous affectons la valeur centrale de la plage à laquelle il appartient. La figure 4.4 illustre les résultats visuels des deux types de quantification. La première remarque que nous pouvons faire à partir de ces résultats est que l'intensité des couleurs entre les deux quantifications n'est pas la même. Quand k_{quanti} est égal à 4, la couleur du mur (à gauche dans l'image) n'est pas la même dans les deux quantifications. Le résultat de k-Moyenne se rapproche le plus de l'image orignale avec des couleurs qui sont plus proches. Le tableau 4.1 présente les valeurs quantitatives de l'EQM. Nous remarquons que k-Moyenne est la méthode qui a des valeurs du EQM inférieures à celles de la méthode de quantification fixe. k-Moyenne est l'algorithme optimal dans le choix des centroïdes par rapport à l'ensemble des valeurs des pixels [78]. Ensuite, le choix de la valeur du k_{quanti} reste important car elle doit s'adapter au type de donnée traitée. Dans l'exemple présenté dans la figure 4.4,

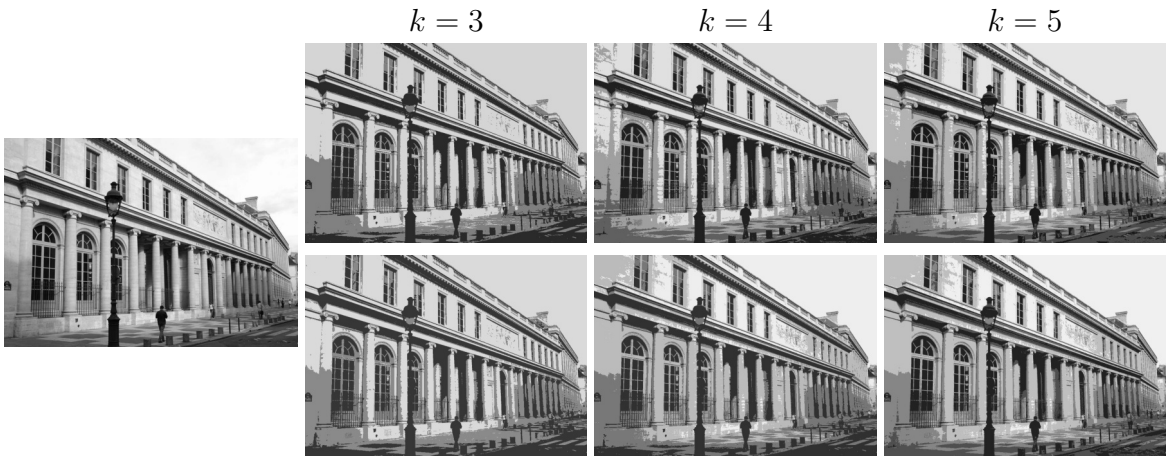


Figure 4.4 – Résultat visuel de la quantification avec deux méthodes ; (Haut) quantification par plages fixes ; (Bas) quantification avec k-Moyenne ; (a) l'image originale ; (b) les images quantifiées avec différentes valeurs de k_{quanti}.

la meilleure valeur de k_{quanti} est 4 avec une confirmation basée sur le calcul de la différence relative (DR) entre les deux types de quantifications via l'équation suivante :

$$DR(a, b) = \frac{|a - b|}{b} \tag{4.6}$$

Tableau 4.1 – Évaluation des méthodes de quantification pour différentes valeurs de k_{quanti}.

k_{quanti}	2	3	**4**	5	6
division en plages (a)	1306.05	601.88	**387.42**	226.61	153.60
k-Moyenne (b)	1230.58	571.50	**318.55**	195.38	134.57
$DR(a, b)$	0.06	0.05	**0.18**	0.14	0.12

Grâce à cette quantification, il y a une plus forte probabilité de trouver des valeurs égales successives. L'algorithme RLE peut alors être directement appliqué sur les pixels de $(J_t)_t$ sans avoir besoin de modifier la fonction d'égalité P.

Afin d'illustrer le résultat de la quantification et les caractéristiques définies précédemment (dans la section 4.2), nous avons généré des données synthétiques qui représentent des pixels temporels. Ces pixels partagent la même évolution temporelle. Nous supposons qu'ils sont tous d'une même catégorie. Pour ce faire, chaque pixel temporel est constitué à partir d'une somme de Gaussiennes avec différents paramètres (moyenne et écart-type). Aussi, nous rajoutons du bruit dans le but d'avoir des variations dans ces pixels temporels. La partie gauche de la figure 4.5 illustre un des pixels temporels synthétiques. Après application de l'algorithme de k-Moyenne sur l'ensemble des valeurs des pixels, la quantification

du pixel temporel présenté dans la partie gauche est illustré dans la partie droite de la figure 4.5. Les valeurs des différentes caractéristiques extraites sont illustrées sur le résultat du pixel quantifié.

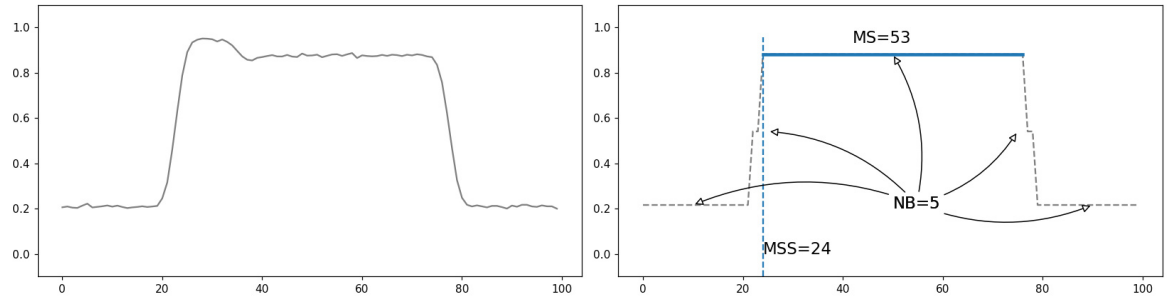


Figure 4.5 – Résultat de la quantification et visualisation des caractéristiques MS, MSS et NB.

L'algorithme RLE est seulement appliqué dans le domaine temporel des pixels. Cela limite la nature des caractéristiques extraites à n'être que temporelles. De plus, les images numériques sont toujours confrontées à différents problèmes (*e.g.*, illumination, bruits). La prise en compte de l'information spatiale autour du pixel permettrait de s'affranchir en partie de ces problèmes tout en étendant les caractéristiques extraites vers une nature spatio-temporelle. Dans la suite, nous allons proposer des relaxations du prédicat P afin que l'information spatiale puisse être prise en considération, amenant à ce que les caractéristiques soient nativement spatio-temporelles.

4.4 Vers la stabilité spatio-temporelle

L'application de l'algorithme RLE est principalement basée sur la notion de « *plages* », calculées sur une séquence ; une « *plage* » est une sous-séquence composée de valeurs identiques répétées successivement. En traitement d'images, les données sont soumises à différentes perturbations. Il y a souvent des images affectées par du bruit, par exemple poivre et sel. De plus, quand une séquence d'images est considérée, nous sommes souvent confrontés à des problèmes d'illumination. Comme notre stratégie repose sur la capacité de compression du RLE, ces bruits potentiels influent sur la qualité des résultats et ainsi sur la qualité de la représentation sous-jacente. Dans le RLE classique, le prédicat P (fonction 4.4) est basé sur une relation d'égalité « stricte ». L'exemple présenté ci-dessous illustre le problème du calcul du RLE sur une donnée bruitée représentée par un pixel $p(x, y)$.

$p(x,y)$	2	5	2	5	2	5	5
$RLE(p)$	1	1	1	1	1	2	

On remarque que les valeurs 2 et 5 se répètent de façon alternative plusieurs fois. Dans ce cas, l'algorithme RLE ne compresse pas efficacement la séquence. Pour résoudre ce pro-

blème, nous allons relaxer l'égalité au niveau du domaine temporel afin de pouvoir vérifier si la valeur o_1 est présente dans un voisinage temporel de la valeur o_2. Une telle relaxation sera aussi appliquée au domaine spatial afin de vérifier si la valeur o_1 est égale à une des valeurs voisines spatialement de o_2. Il est à noter que o_1 et o_2 sont les arguments du prédicat P. La combinaison des deux relaxations temporelle et spatiale en même temps amène à une relaxation spatio-temporelle. De cette façon, l'algorithme RLE ne sera plus un algorithme de compression sans perte mais sera juste un algorithme de compression approximative qui peut absorber différents types de bruit, nous noterons $\widetilde{RLE}$ la version modifiée.

Pour calculer la transformation par l'algorithme RLE, la stratégie consiste ici à calculer des « *plages* » sur la séquence en modifiant la fonction P utilisée pour estimer l'égalité entre des valeurs successives.

Dans ce cas, $P^*(o_1, o_2)$ ne comparera pas directement les deux objets passés en paramètres. o_1 est la valeur du pixel temporel à l'instant t et o_2 sa valeur à l'instant $t+1$. Elle utilisera une fonction intermédiaire $v^*_{w_*}$. $v^*_{w_*}(o_2)$ va sélectionner les objets voisins de o_2 selon le domaine $*$ sur lequel la relaxation est réalisée qui peut être temporelle ($temp$), spatiale ($spatio$) ou spatio-temporelle ($spatio - temp$). v_{w_*} est une fonction qui retourne une liste dont les valeurs dépendent du type de la relaxation. Grâce à cette modification, une relaxation de l'égalité permettra d'être robuste aux bruits car P^* sera valide si o_1 est égal à une des valeurs de la séquence obtenue avec $v^*_{w_*}(o_2)$. Bien entendu, l'objectif est de trouver dans la séquence les plages les plus longues constituant une partition des composants du pixel temporel, cela se fait de manière à rendre $\widetilde{RLE}$ le plus court et avec un maximum MS le plus grand possible. Les relaxations introduites ne permettent plus de traiter le pixel temporel en séquence. Pour illustrer cela, nous reprenons l'exemple précédent et nous relaxons l'égalité sur le domaine temporel avec $w_t = 1$. $v^{temp}_{w_t=1}$ retourne alors une séquence qui ne contient que deux objets. Séquentiellement, deux plages sont obtenues dont la plus longue est égale à 5. Par contre, une plage commençant au deuxième élément du pixel temporel est de longueur 6. Les deux résultats sont illustrés ci-dessous :

$p(x,y)$	2	5	2	5	2	5	5
$\widetilde{RLE}(p)$			5			2	

$p(x,y)$	2	5	2	5	2	5	5
$\widetilde{RLE}(p)$	1			6			

Afin de ne récupérer que le résultat optimal, nous serons donc amené à proposer un algorithme sous-optimal pour le $\widetilde{RLE}$. Pour cela, nous avons implémenté un algorithme glouton afin d'étudier toutes les plages à partir de chaque élément du pixel temporel, c'est-à-dire étudier toutes les plages commençant à une date donnée. Nous notons $L(p_t)$ la longueur de la plage commençant à l'instant t. Nous sélectionnons ensuite la plage la plus longue, notée LR comme *Long Run*, de l'élément en l'instant t (le MS le plus grand).

$$LR(p) = \max_t L(p_t) \quad \text{et} \quad ilr(p) = \min_{t \in [1,T]} \{L(p_t) = LR(p)\} \tag{4.7}$$

où $ilr(p)$ est une fonction qui fournit la position (indice) du point de départ de la plus longue « *plage* » de p. Le $LR(p)$ est conservé comme élément final du $\widetilde{RLE}$.

Une fois que le calcul de LR fixé, nous pouvons en déduire MS, par contre, pour calculer NB, nous appliquons à nouveau ce processus, de manière récursive (en utilisant un paradigme de division et de conquête), à gauche et à droite de la « *plage* » sélectionnée, jusqu'à ce que nous ayons considéré toutes les valeurs temporelles du pixel. Le but de cette opération est de minimiser le nombre de changements dans la description donnée par $\widetilde{RLE}$. La relation fondamentale du processus récursif peut être écrite comme suit :

$$\text{Si } p = (p_t)_{t=1}^{T} \text{ avec } \exists a \in \mathbb{R}^B,\ \forall t \in [1,T]\,,\ p_t = a \text{ alors } \widetilde{RLE}(p) = T \tag{4.8}$$

$$\widetilde{RLE}(p) = \left(\widetilde{RLE}(p_1 \dots p_{ilr(p)-1}), LR(p), \widetilde{RLE}(p_{ilr(p)+LR(p)} \dots p_T)\right) \tag{4.9}$$

Maintenant que le $\widetilde{RLE}$ est optimisé pour maximiser le LR, nous allons évaluer sa complexité algorithmique. Le RLE original a une complexité linéaire par rapport à T le nombre de dates de la séquence ($\mathcal{O}(T)$). Le calcul du LR se fait de manière gloutonne par l'application du $\widetilde{RLE}$ sur le pixel temporel en commençant à chaque date composant la séquence, conduisant à un processus de complexité quadratique ($\mathcal{O}(T^2)$). Concernant l'aspect récursif, l'arbre des appels est comparable à celui sous-jacent à l'algorithme de tri rapide récursif qui est de complexité quasi-linaire ($\mathcal{O}(n \times \log(n))$ où n est le nombre d'éléments à trier : le nombre d'étage de l'arbre est de l'ordre de $log(n)$ et le traitement de chaque étages conduit à effectuer n opérations). En revenant dans le cas du $\widetilde{RLE}$, le traitement de chaque étage conduit à T^2 opérations. La complexité de l'algorithme proposé est ainsi quasi-quadratique par rapport à T ($\mathcal{O}(T^2 \times \log(T^2))$).

Nous définissons maintenant plusieurs façons de considérer qu'une sous-séquence est « constante » et de calculer une « *plage* » approximative commençant à une position donnée. Cela se fera par la modification de la fonction P dans les domaines temporel, spatial et enfin dans les deux domaines simultanément.

Par la suite, nous allons voir comment calculer une « *plage* » approximative en fixant les différentes modifications de la fonction P conduisant aux différentes relaxations. Soit un extrait d'un pixel temporel de la STI $p(x,y) = (p_i)$ où i est inclus dans $[1,T]$.

- **Relaxation temporelle :** la première possibilité que nous proposons consiste à assouplir la contrainte d'égalité sur le domaine temporel. Pour cela, nous considérons une nouvelle définition du prédicat P^{temp} comme :

$$P^{temp}(p_i, p_{i+1}) = \begin{cases} Vrai & \text{si} \quad \exists\, j \in [1, w_t+1]\,, P\Big(p_i, \big(v_{w_t}^{temp}(p_{i+1})\big)_j\Big) \\ Faux & \text{sinon} \end{cases} \tag{4.10}$$

  où $v_{w_t}^{temp}$ retourne une séquence qui contient uniquement des valeurs successives temporelles de p_{i+1}. La fonction P^{temp} va ignorer certains bruits dans le temps lors de la comparaison des éléments consécutifs de la séquence, les valeurs à sauter dépendent de w_t. Par exemple si $w_t = 2$ alors $v_2^{temp}(p_{i+1}) = \{p_{i+1}, p_{i+2}, p_{i+3}\}$.

- **Relaxation spatiale :** lorsque nous relaxons le domaine spatial, la valeur du pixel à l'instant i est comparée à la valeur suivante en $i+1$ et à ses voisins directement connectés. Les voisins sont limités à une fenêtre carrée de taille $w_s \times w_s$. Cela donne une certaine souplesse spatiale pendant la comparaison. Pour ce faire, la fonction $v_{w_s}^{spatio}$ est introduite pour retourner une séquence de pixels voisins selon la taille de la fenêtre w_s. Grâce à cette stratégie, nous pouvons éviter le bruit (poivre et sel) et les problèmes potentiels de recalage des images dans le cas où les images de la STI ont été recalées. Le nouveau prédicat P^{spatio} est défini comme :

$$P^{spatio}(p_i, p_{i+1}) = \begin{cases} Vrai & \text{si} \quad \exists\, j \in [1, w_s^2]\,, P\Big(p_i, \big(v_{w_s}^{spatio}(p_{i+1})\big)_j\Big) \\ Faux & \text{sinon} \end{cases} \tag{4.11}$$

- **Relaxation spatio-temporelle :** une comparaison complètement relaxée est basée sur l'utilisation des deux fonctions $v_{w_s}^{spatio}$ et $v_{w_t}^{temp}$. La nouvelle fonction $P^{spatio-temp}$ compare p_i avec chaque valeur des voisins spatiaux sélectionnés par $v_{w_s}^{spatio}$ et ce pour chaque valeur de la sous-séquence temporelle validée par $v_{w_t}^{temp}$. La fonction $P^{spatio-temp}$ est définie comme :

$$P^{spatio-temp}(p_i, p_{i+1}) = \begin{cases} Vrai & \text{si} \quad \begin{array}{c} \exists\, k \in [1, w_t+1]\,, \\ \exists\, j \in [1, w_s^2] \end{array} \quad P\Big(p_i, \big(v_{w_s}^{spatio}(v_{w_t}^{temp}(p_{i+1})_k)\big)_j\Big) \\ Faux & \text{sinon} \end{cases} \tag{4.12}$$

La figure 4.6 illustre les trois relaxations associées aux trois fonctions P^{temp}, P^{spatio} et $P^{spatio-temp}$. Compte tenu de ces trois prédicats, nous pouvons maintenant calculer la longueur de la plus longue série, à partir de p_i. La longueur de la « plage » la plus longue (LR) peut alors être calculée en utilisant la fonction compteur c comme :

$$P^*(p_i, p_j) = P^*(p_i, p_{i+1}) \times \prod_{k=i+1}^{j-1} P^*(p_k \leftarrow p_i, p_{k+1}) \tag{4.13}$$

$$c(P^*(p_i, p_j)) = \begin{cases} j-i+1 & \text{si} \quad P^*(p_i, p_j) \\ 0 & \text{sinon} \end{cases} \tag{4.14}$$

où $P^\star$ est l'un des prédicats relaxés mentionnés ci-avant.

En considérant ces différentes stratégies de relaxation, nous pouvons obtenir plusieurs approximations du RLE pour un pixel temporel p, notées $\widetilde{RLE}_{temp}(p)$ en considérant la relaxation temporelle, $\widetilde{RLE}_{spatio}(p)$ en considérant la relaxation spatiale, et $\widetilde{RLE}_{spatio-temp}(p)$ en considérant la relaxation spatio-temporelle.

En analyse de vidéo, certaines méthodes sont conçues pour résumer une vidéo en une ou quelques images. Nous allons voir par la suite comment combiner les trois caractéristiques de stabilité que nous avons définies pour réaliser un résumé de la STI en une seule image en fausses couleurs.

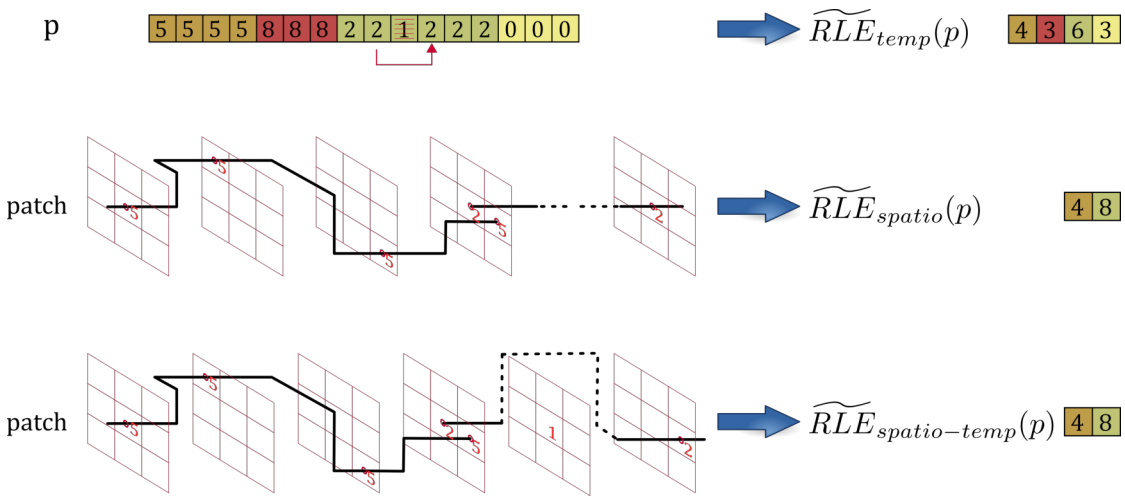


Figure 4.6 – Illustration du calcul du $\widetilde{RLE}$ avec les différentes approximations. (En haut) égalité avec la relaxation temporelle ; (Centre) égalité avec la relaxation spatiale ; (En bas) égalité avec la relaxation spatio-temporelle.

4.5 Résumé 2D d'une séquence

En analyse de STI, diverses méthodes se focalisent sur la génération d'un résumé qui permet de représenter la STI par une ou quelques images représentatives. En télédétection, certaines méthodes analysent les motifs les plus répétés dans les pixels temporels [85, 75] mais ne permettent pas de les résumer. En analyse de vidéo, le mot « résumé » ne sous-entend pas forcément d'avoir une seule image qui résume la vidéo mais de créer un court-métrage de la vidéo originale, tout en préservant les éléments principaux du contenu [96].

Dans notre cas, nous proposons d'utiliser les trois caractéristiques extraites par la méthode proposée, et selon les différentes stratégies d'égalité « stricte » ou relaxées, pour résumer les STI, en les combinant dans une image en fausses couleurs. Ce résumé est noté alors TS pour le RLE classique (sans relaxation) et pour chaque $\widetilde{RLE}$ approximatif, Nous notons TS_{temp}, TS_{spatio} et $TS_{spatio-temp}$ (pour les différents $\widetilde{RLE}$) les images de synthèse qui en résultent. La composition en couleurs de l'image résultante est la suivante : MS dans le canal rouge (R) ; NB dans le canal vert (V) et MSS dans le canal bleu (B).

La figure 4.7 présente quatre images d'une vidéo synthétique avec les différentes caractéristiques extraites et leurs compositions en une image couleur pour former le résumé TS. La vidéo synthétique contient un carré qui bouge de gauche à droite. La première remarque concerne l'arrière plan qui reste stable pendant la durée de la vidéo. Celui-ci est rouge dans TS car MS capte bien cette stabilité qui commence au début de la vidéo, caractérisé par MSS, avec aucun changement comme représenté dans NB. Par contre le carré est plus stable dans les extrémités du mouvement à cause du recouvrement de ces parties entre les images. Cela est représenté dans TS en dégradé de vert au jaune qui commence du centre

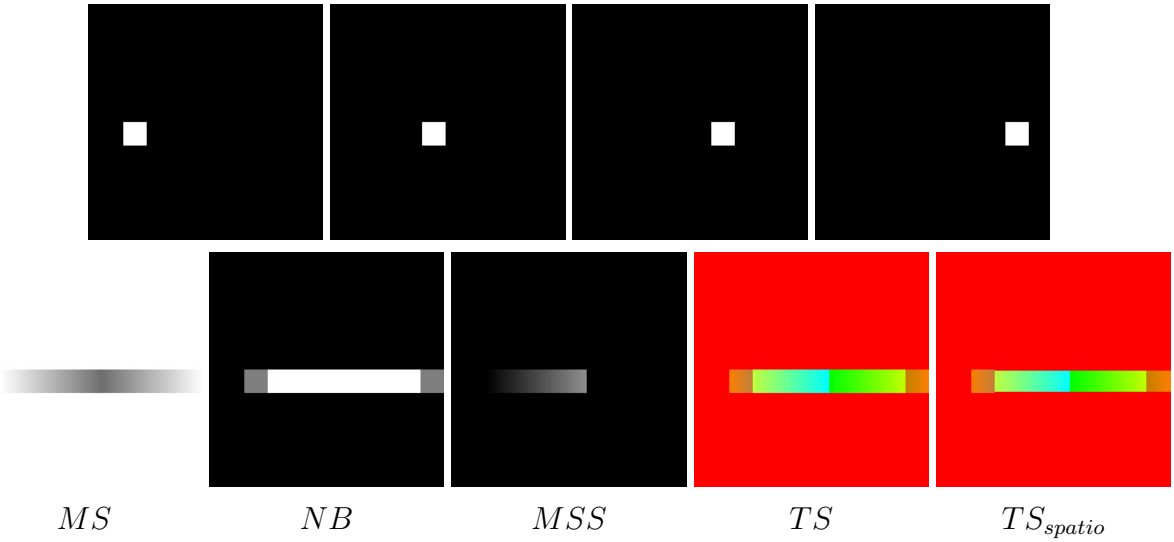


Figure 4.7 – Illustration de quelques images d'une vidéo synthétique avec les différentes caractéristiques extraites et les résumés associés.

allant aux extrémités du chemin de déplacement. La couleur bleutée dans TS est due au MSS qui montre que MS commence très tard dans la moitié du chemin (moitié gauche) du carré. Le résumé TS_{spatio} présente le résultat obtenu quand l'égalité est relaxée sur le domaine spatial avec w_s égal à 3. Nous observons que le carré est visible dans le point du départ (par rapport au chemin de déplacement du carré). La relaxation spatiale détecte que les bords sont stables au court du temps. Cela est présenté comme une érosion de la trajectoire du carré. TS_{temp} et $TS_{spatio-temp}$ sont respectivement identiques à TS et TS_{spatio} dans cet exemple.

Par l'intermédiaire de cette image de résumé, au lieu d'analyser toute la séquence d'images, il est possible d'analyser une seule image qui résume toute la STI.

4.6 Étude expérimentale

Dans cette partie, une étude expérimentale est menée. Tout d'abord, les résultats du résumé de la méthode proposée seront présentés dans la section 4.6.1. Ensuite, la section 4.6.2 présente les résultats quantitatifs lors de l'implication des caractéristiques de stabilité dans un problème de classification. Dans ces deux sections, les résultats présentés concernent les deux cadres applicatifs. La première concerne un problème de télédétection qui est l'analyse de la couverture urbaine à partir de $STIS$ et la deuxième est celle de classification de vidéos en fonction de la présence de violence.

4.6.1 Visualisation et interprétation du résumé de stabilité

Le cœur de la méthode se base sur la répétition des valeurs successives dans le temps. Dans ce contexte, pour simplifier le problème, notre méthode est appliquée sur des données scalaires seulement. Pour cela, dans le cadre de l'étude relative à l'analyse de la couverture des sols, nous avons choisi de considérer un indice spectral de télédétection qui est le $NDVI$. Ce dernier est largement utilisé dans les études de télédétection pour analyser la couverture des sols à partir des $STIS$, car il est sensible à l'état de la végétation. Le calcul du $NDVI$ est basé sur les bandes Nir et R, conduisant à une $STIS$ de $NDVI$, notée $(I_t^{NDVI})_{t=1}^T$:

$$NDVI(I_t) = \frac{I_t^{Nir} - I_t^R}{I_t^{Nir} + I_t^R}, \quad \forall t \in [1, T] \tag{4.15}$$

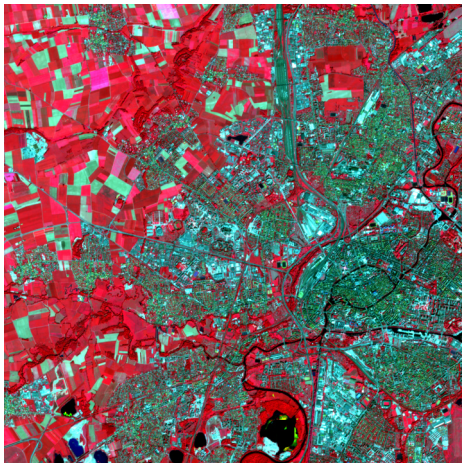
Concernant les vidéos, ces données seront traitées en niveaux de gris, conduisant à une STI en niveaux de gris, notée $(I_t^{gris})_{t=1}^T$. Ensuite, selon notre stratégie de quantification présentée dans la section 4.3, les STI sont quantifiées, donnant $(J_t^{NDVI})_{t=1}^T$ et $(J_t^{gris})_{t=1}^T$. Cependant, le choix du k_{quanti} est une étape cruciale car la quantité d'information qui sera gardée dépend de ce paramètre. Dans ce contexte, le tableau 4.1 présenté précédemment (page 48) illustre les résultats de quantification avec différentes valeurs de k_{quanti}. Pour cela, nous avons utilisé la même procédure pour choisir k_{quanti}. En se basant sur la différence relative entre les valeurs de l'EQM des deux méthodes de quantification, la meilleure valeur du k_{quanti} est 4 et nous l'utilisons dans le reste des expérimentations.

Les caractéristiques proposées sont ensuite calculées à partir des données quantifiées. Enfin, les caractéristiques extraites sont combinées en des images en fausses couleurs, comme expliqué dans la section 4.5, qui résume les STI de T images en une image unique. Concernant les relaxations temporelles et spatiales associées respectivement aux équations 4.10 et 4.11, nous fixons w_t à 1 et w_s à 3. De cette façon, la relaxation temporelle va pouvoir sauter au maximum une image et la relaxation spatiale vérifie si l'égalité est validée entre un pixel à l'instant t avec le pixel à l'instant $t+1$ ou avec un parmi ses huit voisins.

Dans la suite, comme les données utilisées dans nos expérimentations ne sont pas du même type, nous allons présenter les résultats obtenus séparément. Les résultats sur les données satellitaires sont présentés dans un premier temps puis nous passerons aux résultats obtenus sur les vidéos.

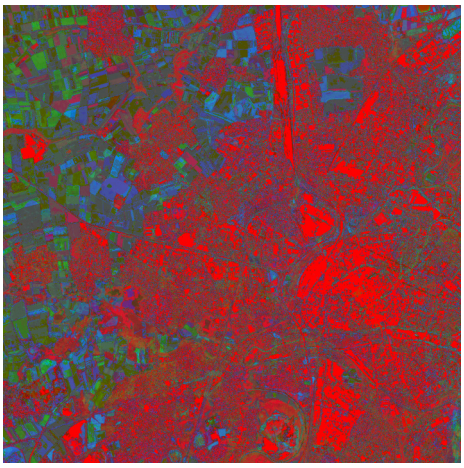
4.6.1.1 Résumé des $STIS$

Nous rappelons que les $STIS$ présentent une scène qui évolue dans le temps. Nous considérons aussi, à titre de comparaison, une stratégie naïve qui permet de résumer la série et sera considérée comme la base de référence. Cette stratégie consiste à calculer, pour chaque pixel du $STIS$ original I_t^{NDVI}, la moyenne (dans le temps) des valeurs des pixels

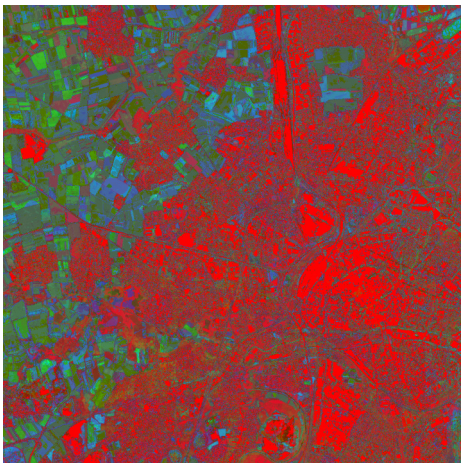


(a) 2017-08-26

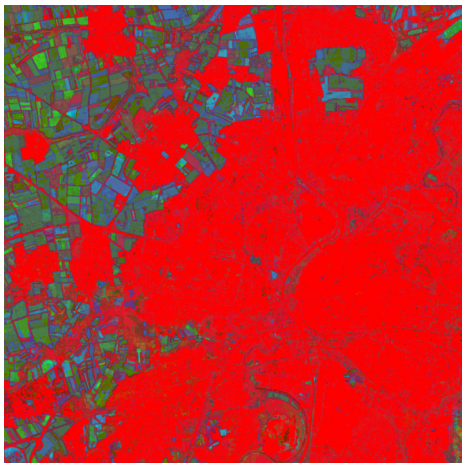
(b) $\overline{p^{NDVI}}$



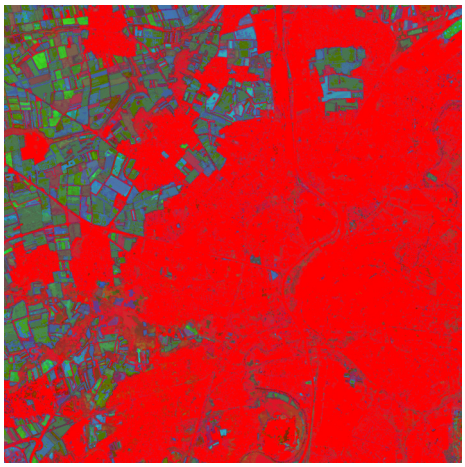
(c) TS



(d) TS_t



(e) TS_s



(f) $TS_{spatio-temp}$

Figure 4.8 – Illustration des données et des résumés correspondant à la zone géographique de Strasbourg.

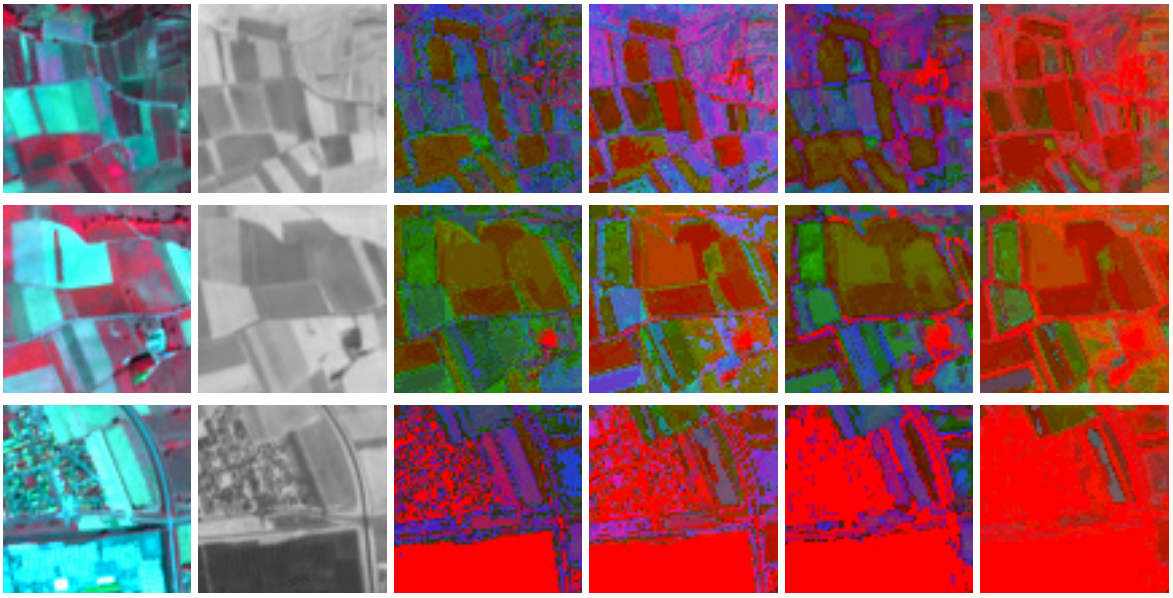
temporels $(p_t)_{t=1}^{T}$ pour obtenir une seule valeur scalaire de $NDVI$ par pixel temporel. Le résultat de cette stratégie naïve est noté $\overline{p^{NDVI}}$.

La figure 4.8 présente une image de la série à la date de 26/08/2017 sur la zone géographique de Strasbourg ainsi que l'image moyenne $\overline{p^{NDVI}}$ et les résumés proposés avec / sans les différentes relaxations. Nous remarquons qu'avec $\overline{p^{NDVI}}$, l'intensité des pixels dans les zones urbaines est très sombre car l'intensité du NDVI est grande dans les zones de végétation et faible dans les autres zones. Dans les zones agricoles, nous visualisons que nous avons au moins quatre niveaux de gris perceptibles, ce qui signifie la présence de différentes pratiques de gestion agricole. Par contre, $\overline{p^{NDVI}}$ a ses limites car la distinction entre les deux types de classes, zones naturelles ou zones artificielles, est difficile dans certaines zones péri-urbaines.

Les images obtenues avec la méthode proposée permettent une meilleure visualisation et discrimination de nombreuses classes thématiques. Les couleurs des pixels sont liées à l'évolution des pixels temporels dans le temps et peuvent aider l'expert à mieux interpréter la scène. De telles observations peuvent facilement permettre à un utilisateur, comme un géographe, de mieux interpréter les territoires observés en ne considérant qu'une seule image au lieu d'une série entière. Par exemple, les zones artificialisées sont en couleur rouge. Une telle image capte des phénomènes spatio-temporels et les résume en trois valeurs pour chaque pixel seulement. À première vue, les zones les plus stables sont rapidement identifiées et ce surtout dans les zones urbaines. Elles sont rouges car la région reste stable pendant une longue période (MS élevé) et ce dès le début de l'année (faible MSS) avec peu de changements (NB faible). Les zones péri-urbaines sont plus difficiles à être localisées à cause de la présence de végétation (jardin, arbre). Ces dernières sont représentées par un mélange de rouge et d'autres couleurs (vert, bleu ou violet). Le vert signifie que la région change beaucoup au fil du temps avec les petits MS et MSS. Le bleu exprime une faible stabilité MS mais qui débute très tard (MSS élevé). Et le violet présente une équivalence entre les valeurs de stabilité MS et son début MSS. Grâce à la relaxation spatiale, ces zones sont mieux distinguées. Le reste des zones représente les régions agricoles. Plusieurs couleurs sont observées à cause des différentes pratiques agricoles et de leurs gestions (saisons, fauchage, pâturage).

Pour une meilleure visualisation, la figure 4.9 présente un zoom sur trois régions géographiques différentes. Les deux premières lignes sont centrées sur des zones agricoles et la troisième sur une zone péri-urbaine. La colonne (a) présente l'image originale de la $STIS$ à la même date du 26/08/2017, (b) présente les résultats de $\overline{p^{NDVI}}$ et (c), (d), (e), (f) sont les résultats des quatre résumés proposés avec / sans les différentes relaxations.

En comparant les résultats obtenus avec les différentes relaxations, nous remarquons visuellement qu'il n'y a que de petites différences entre TS et TS_{temp} où certaines parcelles sont devenues plus homogènes. Par contre dans la région urbaine, nous remarquons qu'il y a plus de zones stables car nous échappons aux pixels correspondant à la végétation des jardins ou des plantes sauvages. Par contre avec TS_{spatio} et $TS_{spatio-temp}$, les routes, les chemins et les délimitations de parcelles sont plus facilement visibles entre les champs agricoles et dans les environnements urbains. Les relaxations ont amélioré la qualité visuelle



(a) (b) (c) (d) (e) (f)

Figure 4.9 – Illustration des données et des résultats des résumés dans trois zones géographiques différentes : (a) Image du $STIS$ à la date du 2017-08-26 ; (b) Résultat de la moyenne des $STIS$ de $NDVI$ $\overline{p^{NDVI}}$; (c),(d),(e) et (f) Résultats de l'approche proposée avec les différentes relaxations, liées respectivement à TS, TS_{temp}, TS_{spatio} et $TS_{spatio-temp}$.

des résumés par rapport à l'égalité stricte du RLE qui donne l'image TS. Les délimitations de routes sont censées avoir une grande stabilité temporelle mais ce n'est pas vraiment observé car elles sont peu visibles avec TS. Cela est dû à des problèmes de recalage et de discrétisation du signal qui approximent la réalité à une résolution de 10 mètres. Par exemple, la route dans une image apparait avec un léger décalage dans l'image suivante car les pixels ne sont pas centrés de façon exacte. Par contre, lorsque le domaine spatial est relaxé avec TS_{spatio} et $TS_{spatio-temp}$, nous avons la possibilité de limiter ce type de problèmes et nous optimisons les résultats en réduisant le bruit (*e.g.*, type poivre et sel).

4.6.1.2 Résumés des vidéos

Contrairement aux données satellitaires, il n'est pas pertinent dans le cadre de l'application vidéo de calculer l'image moyenne car elle n'aura aucun sens à cause du mouvement de la caméra ou des objets qu'il y a dans la scène. La figure 4.10 présente des résumés de deux vidéos, une violente et une non violente de la base *Movies Fights* [91]. Pour chaque vidéo, quatre images sont présentées afin d'avoir une idée du mouvement qu'il y a dans la scène. Les couleurs sont interprétées de la même façon qu'avec les $STIS$. La vidéo non violente (vidéo du haut dans la figure 4.10) contient une personne qui fait un geste de « salutation » avec sa main droite. La vidéo violente (vidéo du bas dans la figure 4.10) présente un combat entre Bruce Lee et Chuck Norris. Le mouvement est ici un coup de jambe porté par Bruce

Lee. Visuellement, l'arrière plan est bien rouge dans les deux vidéos. Cela est dû à une valeur élevée de MS (grande stabilité) avec un faible MSS (la stabilité commence très tôt). Les régions qui changent sont visibles en couleur. Une autre perception est que le contour des objets en mouvement est bien visible et le contenu des objets est rouge. Cependant, ce phénomène n'est visible qu'avec les objets qui ne bougent pas rapidement. Par exemple, le corps de Chuck Norris présente bien ce phénomène mais pas celui de Bruce Lee à cause de sa rapidité de mouvement.

Les résumés présentés ici sont relativement homogènes à cause de la caméra fixe et du faible mouvement dans la scène. Par contre, la première remarque est qu'il n'y a pas de différence visuelle entre les résumés de TS et TS_{temp}. Mais une fois la relaxation spatiale introduite, nous remarquons que plusieurs bruits sont supprimés. Ces bruits sont dus à la luminosité de la scène au cours du tournage.

Les résumés peuvent avoir un sens quand la caméra ou la scène ne contient pas beaucoup de variations. Par contre si nous prenons une vidéo de la base *Crowd Violence* [49],

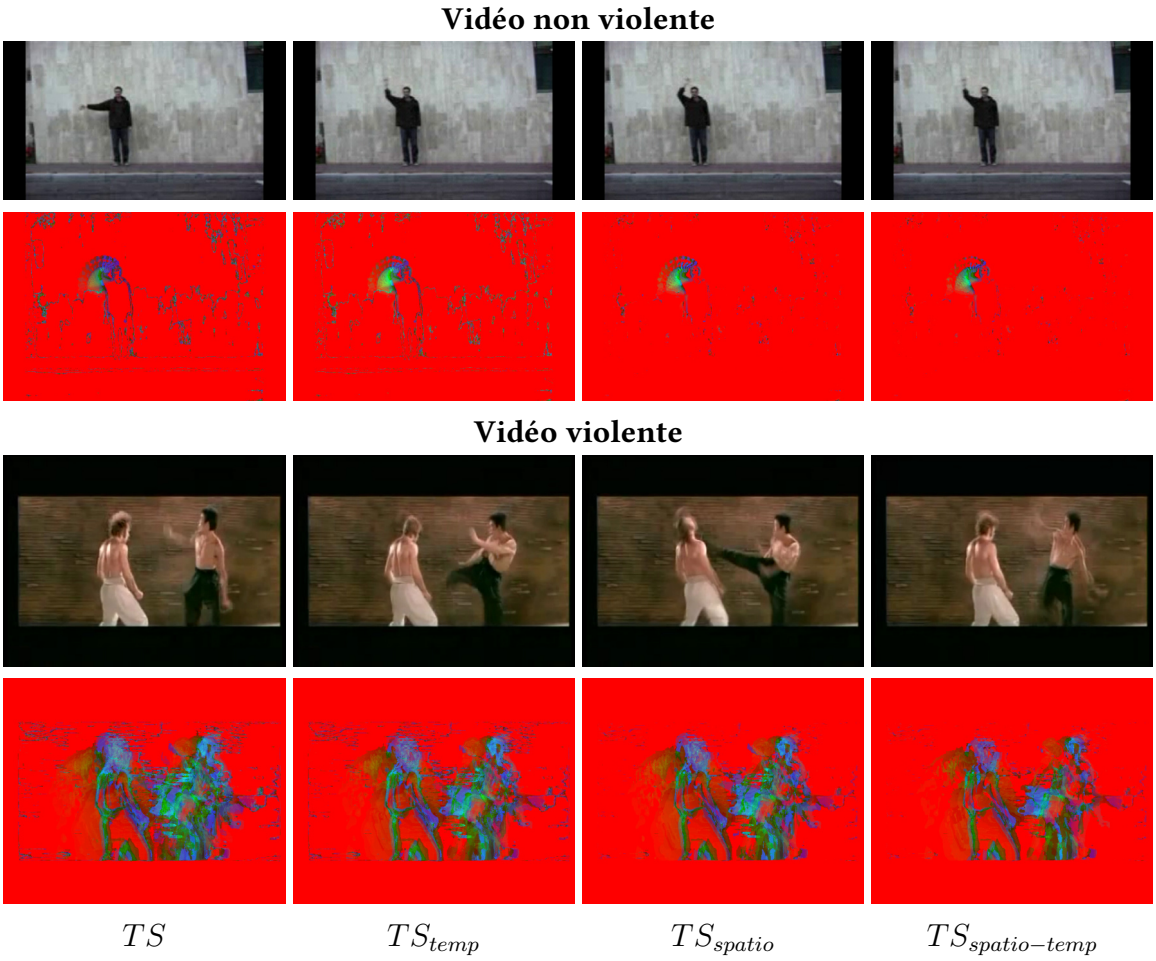


Figure 4.10 – Illustration de deux vidéos avec les différents résumés associés aux différentes relaxations. Les vidéos sont issues de la base *Movies Fights* [91].

toutes les vidéos de cette base contiennent un fort mouvement qui est dû au mouvement des objets dans la scène ou même causé par le mouvement de la caméra.

La figure 4.11 présente deux vidéos, une violente et une non violente, de cette base avec les différents résumés. Les résultats obtenus sont visuellement difficilement interprétables et cela même avec les différentes relaxations. Les seules informations que nous pouvons interpréter sont la date et l'heure en texte incrusté et dans d'autres vidéos les bandes noires qui sont généralement en haut et en bas pour cadrer la vidéo.

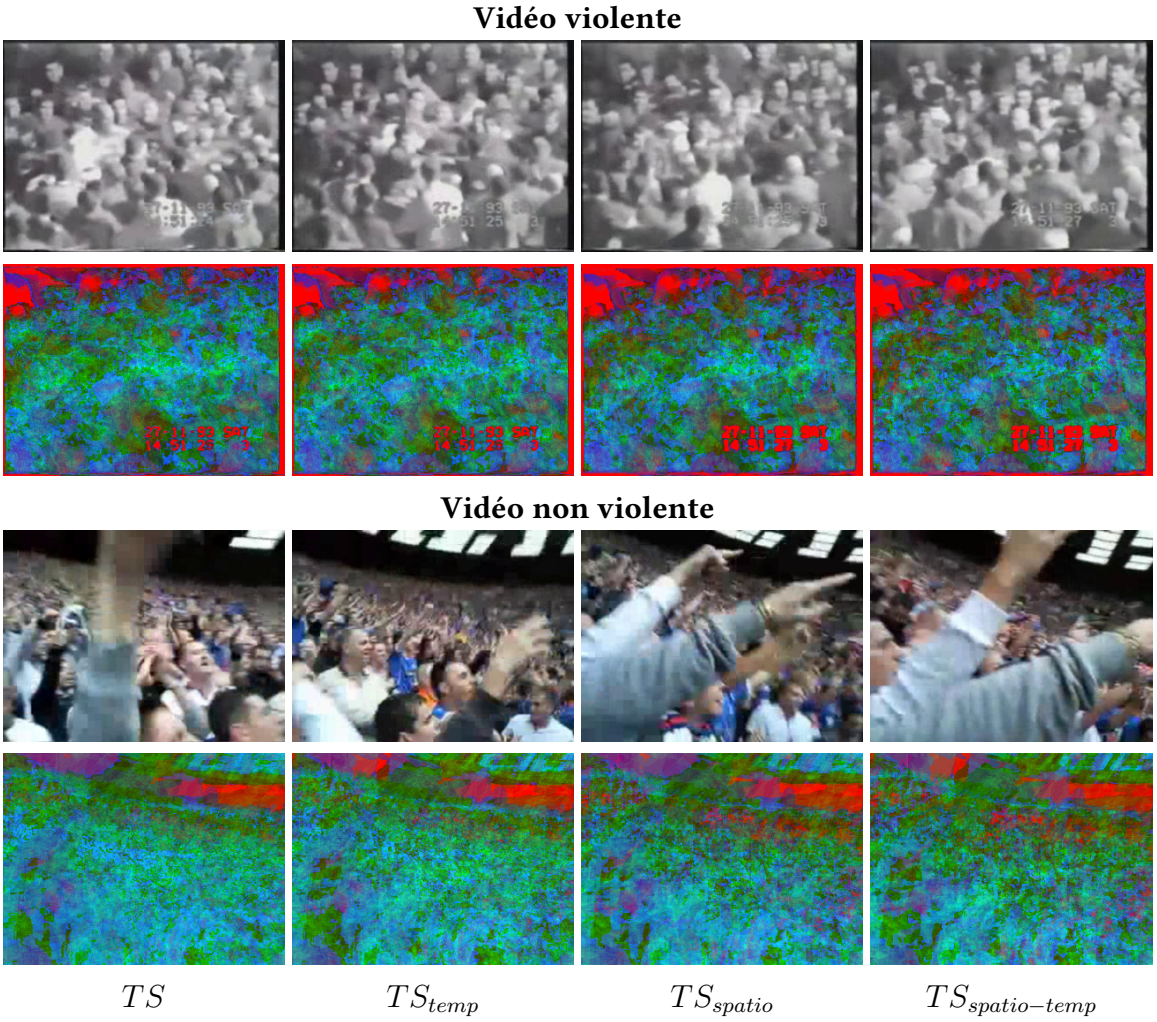


Figure 4.11 – Illustration de deux vidéos issues de *Crowd Violence* [49] avec les différents résumés associés aux différentes relaxations.

4.6.2 Classification des STI

Dans la section 4.6.1, nous avons présenté les résumés construits avec les différentes caractéristiques de stabilité proposées avec/sans relaxations. Nous allons maintenant les employer dans une tâche de classification. Les données satellitaires sont utilisées pour analyser la tache urbaine et les données vidéos pour reconnaître la violence. Les méthodes de classification utilisées sont adaptées à chacune des problématiques.

4.6.2.1 Classification des $STIS$

Dans le contexte de l'analyse de la couverture urbaine, nous allons analyser les performances des trois caractéristiques de stabilité que sont MS, MSS et NB. Dans cette application chaque pixel sera classé individuellement selon deux classes (zone urbaine vs. zone non urbaine). Afin de réaliser une étude comparative, nous considérons également deux stratégies où les pixels à classer sont caractérisés par les trois caractéristiques de stabilité :

1. les pixels temporels $p = (p_t)_{t=1}^{T}$ caractérisés par le $NDVI$, noté p^{NDVI} (50 valeurs des dates, $T = 50$) ;
2. les pixels sont caractérisés par la moyenne de chaque série temporelle, notée $\overline{p^{NDVI}}$ (1 valeur) ;
3. les pixels sont caractérisés par les trois caractéristiques de stabilité, notées TS ou TS_* où $*$ désigne le domaine de relaxation (3 valeurs) ;
4. les pixels temporels sont combinés avec les caractéristiques de stabilité temporelles ($T + 3 = 53$ valeurs).

Enfin, l'influence des différentes stratégies de relaxation introduites précédemment sera également étudiée.

Les algorithmes sélectionnés ont besoin d'un ensemble d'apprentissage et d'un ensemble de test. Dans notre cas, nous utilisons les deux zones sélectionnées comme deux ensembles et nous menons deux expérimentations. La première expérience (Exp 1) consiste à faire l'apprentissage sur la zone de Strasbourg et à tester sur la zone de Mulhouse et dans la deuxième expérience (Exp 2), l'apprentissage est fait sur la zone de Mulhouse puis nous testons sur la zone de Strasbourg.

Les méthodes de classification utilisées dépendent de la taille d'entrée des caractéristiques. Si celle-ci est inférieure à trois (≤ 3), nous utilisons un simple arbre de décision. Sinon une forêt aléatoire est utilisée. L'algorithme d'apprentissage de l'arbre de décision utilisé est le « C4.5 ». Le nombre d'arbres pour la forêt est de 30 où chacun est construit en utilisant le critère de division « Gini » qui permet d'avoir les nœuds les plus purs possible. En plus de ces deux méthodes de classification, nous nous comparons à une méthode basée sur l'apprentissage profond où la convolution n'est appliquée que sur le domaine temporel. La méthode choisie est $TempCNN$ [98].

Le tableau 4.2 présente les taux de classification globaux au niveau pixel avec les différentes méthodes et caractéristiques utilisées sur les zones de test. Tout d'abord, nous

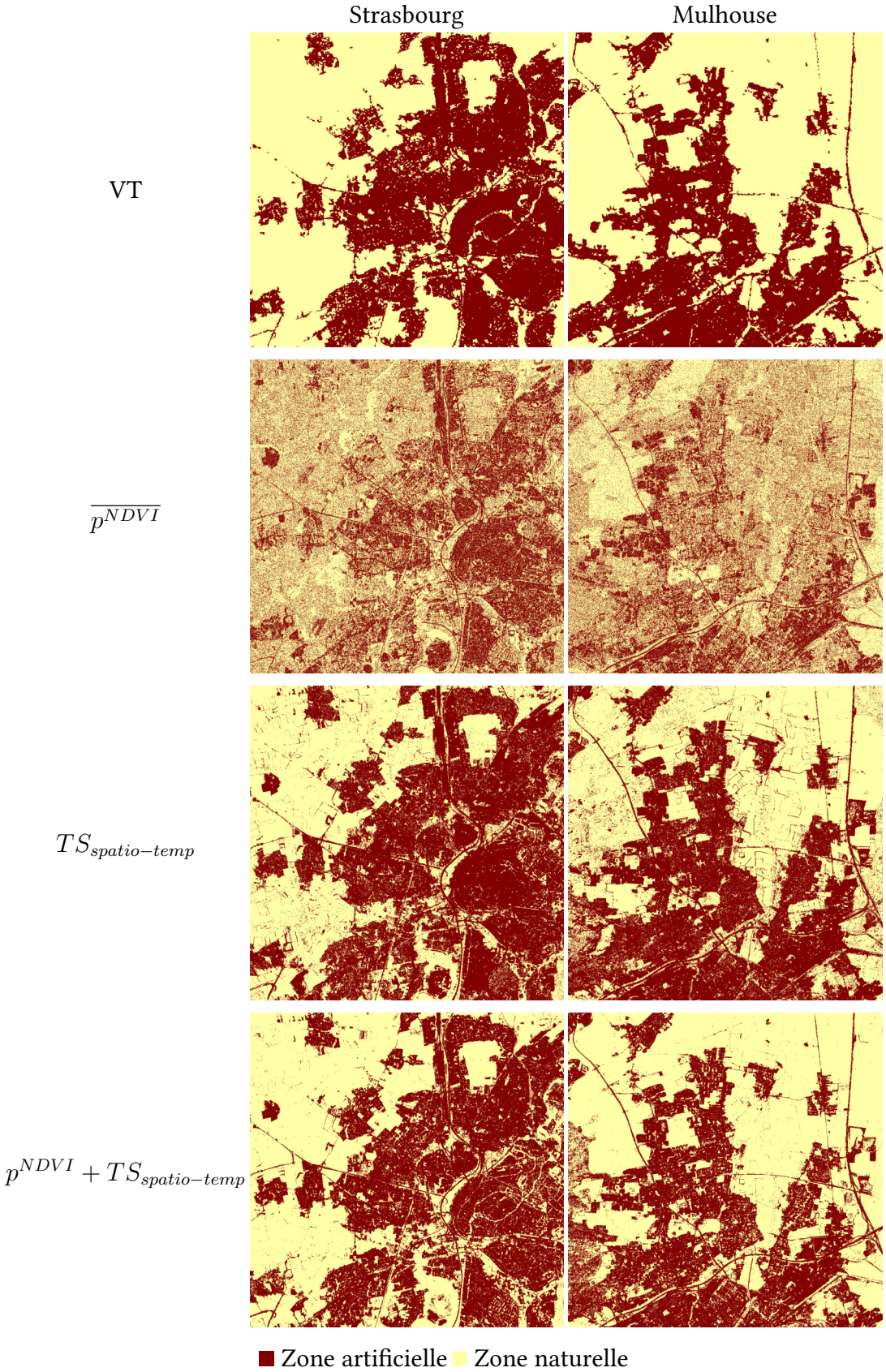


Figure 4.12 – Illustration des VT et des résultats de classification de la tache urbaine obtenus des deux zones géographiques.

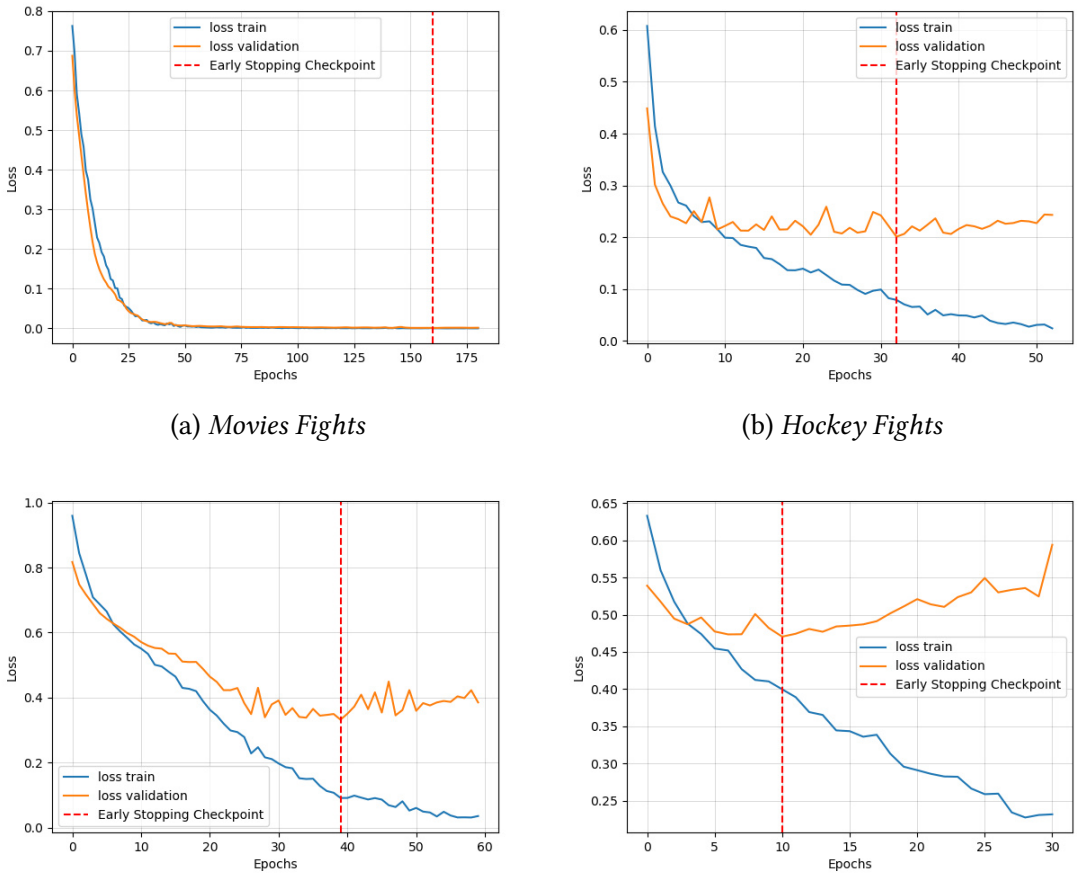
Tableau 4.2 – Résultats quantitatifs dans le cadre de l'application relative à l'analyse de la couverture urbaine (taux de classification global). En gras sont notés les meilleurs résultats.

Classificateur	**Caractéristiques**	**Nb caractéristiques**	**Exp. 1**	**Exp. 2**
Forêt aléatoire	p^{NDVI}	50	84.56	84.02
Arbre de décision	$\overline{p^{NDVI}}$	1	66.40	67.49
CNN	$TempCNN$ [98]	50	84.26	86.05
Arbre de décision	TS	3	72.40	71.58
Arbre de décision	TS_{temp}	3	76.01	79.03
Arbre de décision	TS_{spatio}	3	83.17	85.18
Arbre de décision	$TS_{spatio-temp}$	3	83.33	85.44
Forêt aléatoire	p^{NDVI}+TS	53	84.58	84.15
Forêt aléatoire	p^{NDVI}+TS_{temp}	53	84.56	84.07
Forêt aléatoire	p^{NDVI}+TS_{spatio}	53	**85.85**	**86.23**
Forêt aléatoire	p^{NDVI}+$TS_{spatio-temp}$	53	**85.65**	**86.65**

pouvons remarquer que l'utilisation de p^{NDVI} donne de bien meilleurs résultats que lorsque seule la valeur moyenne $\overline{p^{NDVI}}$ est utilisée. Nous pouvons également constater que l'utilisation des caractéristiques de stabilité que nous avons extraites de la série temporelle donne des résultats encourageants même s'ils sont inférieurs à ceux obtenus en utilisant l'ensemble des p^{NDVI}. Nous constatons en outre l'amélioration apportée par les processus de relaxation proposés. Les caractéristiques spatio-temporelles qui en résultent permettent d'omettre certaines valeurs aberrantes ponctuelles ou d'éliminer les problèmes de recalage que nous avons constatés sur les différentes limites géographiques. La combinaison des caractéristiques de stabilité avec les pixels temporels p^{NDVI} permet d'augmenter les scores de précision. Les caractéristiques TS_{spatio} et $TS_{spatio-temp}$ donnent des résultats nettement plus élevés. Nous pouvons également remarquer que la méthode $TempCNN$ [98], malgré une phase d'apprentissage plus longue, ne présente pas de meilleurs résultats que les nôtres basées sur les p^{NDVI}+TS_{spatio} ou les p^{NDVI}+$TS_{spatio-temp}$ dans cette étude thématique. Dans l'ensemble, le taux d'erreur est réduit entre l'utilisation des caractéristiques de stabilité TS et la séquence p^{NDVI} d'environ 10%. La figure 4.12 illustre différents résultats visuels de la classification lors des deux expériences qui sont : $\overline{p^{NDVI}}$, $TS_{spatio-temp}$ et en combinant les caractéristiques de p^{NDVI} avec $TS_{spatio-temp}$.

4.6.2.2 Classification des vidéos

La classification des vidéos sera basée sur l'utilisation de réseaux de neurones profonds convolutifs. Cependant, l'entrée ne sera pas la vidéo mais un des différents résumés de stabilité. La validation de nos expérimentations est basée sur le protocole de validation croisée



(a) *Movies Fights*

(b) *Hockey Fights*

(c) *Crowd Violence*

(d) *RWF2000*

Figure 4.13 – Courbes de pertes obtenues lors de l'entraînement du modèle sur les caractéristiques de stabilité $TS_{spatio-temp}$ de chacune des bases de vidéos.

à cinq sous-ensembles (5 *folds*). Pour chaque sous-ensemble, nous divisons l'ensemble de données en trois ensembles qui seront utilisés pour l'entraînement, la validation et le test. Les tailles de ces ensembles sont respectivement 60%, 20% et 20% du nombre total de vidéos. Le modèle est ensuite entraîné et est testé cinq fois et nous donnons la précision moyenne. Comme l'ensemble de test est déjà fourni pour la base *RWF2000*, nous extrayons juste un ensemble de validation à partir de celui de l'entraînement. Le modèle est donc entraîné une seule fois dans le cas de cette base.

L'architecture choisie est SqueezeNet [58]. Ce modèle atteint le même niveau de précision que AlexNet quand le modèle est testé sur la base ImageNet sachant qu'il a 50 fois moins de paramètres que AlexNet. Pour des raisons de quantité de données, nous utilisons les poids appris sur ImageNet et nous les raffinons afin qu'ils s'adaptent à notre problème. L'entraînement du modèle se fait en utilisant *Adam* comme optimiseur avec un pas d'apprentissage de 10^{-5} et nous gardons les valeurs par défaut pour le reste des paramètres ($\beta_1 = 0.9, \beta_2 = 0.999$ et $\epsilon = 10^{-8}$). La fonction de perte utilisée est la *cross entropy*. Nous

Tableau 4.3 – Taux de classification obtenus sur la classification des vidéos (les meilleurs résultats sont en gras).

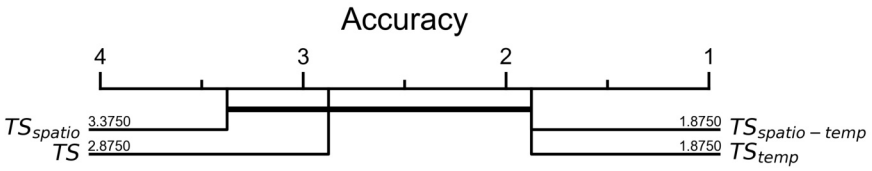
k_{quanti}	**Caractéristique**	*RWF2000*	*Movies fights*	*Hockey fights*	*Crowd Violence*
4	TS	82.5	**97.5**	88.6	80.0
	TS_{temp}	**82.7**	**97.5**	89.9	83.3
	TS_{spatio}	81.5	**97.5**	88.2	80.8
	$TS_{spatio-temp}$	81.7	**97.5**	**91.1**	**84.5**

Tableau 4.4 – Taux de classification des vidéos obtenus avec les méthodes de l'état-de-l'art (les meilleurs résultats sont en gras).

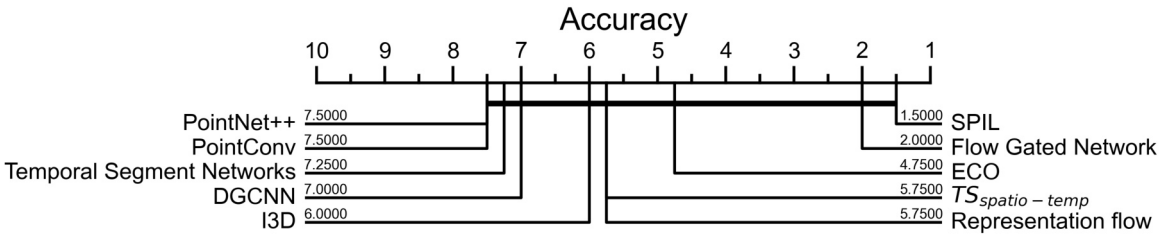
Méthodes	*RWF2000*	*Movies fights*	*Hockey fights*	*Crowd Violence*
CNN 3D				
Temporal Segment Networks [139]	81.5	94.2	91.5	81.5
$I3D$ [18]	83.4	95.8	93.4	83.4
Representation flow [140]	85.3	**97.3**	92.5	**85.9**
Flow Gated Network [22]	**87.3**	n/a	**98.0**	88.8
ECO [153]	83.7	96.3	94.0	84.7
Nuage de points				
$PointNet++$ [102]	78.2	89.2	89.7	89.2
$PointConv$ [148]	76.8	91.3	89.7	89.2
$DGCNN$ [141]	80.6	92.6	90.2	87.4
$SPIL$ [128]	**89.3**	**98.5**	**96.8**	**94.5**

fixons une taille de lot à 32 et l'apprentissage est arrêté avec la technique d'arrêt précoce (*early stopping*) avec un nombre de patience de 10.

La figure 4.13 illustre quatre courbes de pertes quand le modèle est entraîné sur les caractéristiques de stabilité avec une relaxation spatio-temporelle $TS_{spatio-temp}$. Visuellement, le modèle arrive à apprendre car l'erreur diminue au fil des *epochs*. En particulier, ce phénomène est marqué pour la base *Movies Fights* où l'entraînement ne s'arrête qu'après 160 *epochs* avec une erreur très petite. Tandis que pour les autres bases, l'entraînement ne dépasse pas les 40 *epochs*. Par rapport à l'erreur, le modèle arrive à faire réduire l'erreur jusqu'à 0.2 pour la base *Hockey Fights* mais il n'arrive pas à généraliser pour *Crowd Violence* et *RWF2000* car les erreurs pour ces bases ne baissent pas, respectivement en dessous de 0.35 et 0.4.



(a) Comparaison des résultats de notre méthode lors $k_{quanti} = 4$



(b) Comparaison des résultats de notre méthode avec les méthodes de l'état-de-l'art

Figure 4.14 – Diagrammes de différence critique obtenus sur l'ensemble des bases pour la classification de vidéos.

Le tableau 4.3 présente les résultats obtenus quand les valeurs de la STI sont quantifiées avec $k_{Quanti} = 4$. Nous remarquons que ces premiers résultats quantitatifs sont tous acceptables par rapport aux courbes de pertes obtenues et s'adaptent aux difficultés des différentes bases. Afin de faire une comparaison avec les méthodes de l'état-de-l'art, le tableau 4.4 présente les résultats des méthodes basées sur les réseaux 3D et les nuages de points. Les résultats de notre méthode sur la base *Movies Fights* surpassent ceux des modèles 3D et se classent en deuxième position pour les méthodes de nuage de points. Par contre, les résultats obtenus sur les bases *Hockey Fights* et *Crowd Violence* sont en dernière position. Cela peut être dû à la complexité des vidéos et à une quantification trop dure dans notre approche.

Afin d'avoir un classement plus général intégrant les résultats obtenus sur les quatre bases, comparant la méthode proposée et les méthodes de l'état-de-l'art, nous utilisons un diagramme de différence critique [37]. La figure 4.14.a présente le diagramme obtenu uniquement sur les résultats de notre méthode de stabilité quand k_{quanti} est égale à 4. Le diagramme indique que $TS_{spatio-temp}$ est à la première position suivie par TS_{temp}. Par la suite, nous ne gardons que les scores de $TS_{spatio-temp}$ et nous étudions le classement de notre méthode par rapport aux méthodes de l'état-de-l'art. La figure 4.14.b illustre le diagramme obtenu. D'après ce nouveau diagramme, $TS_{spatio-temp}$ se classe en quatrième position derrière trois méthodes de l'état-de-l'art. Ces méthodes sont $SPIL$, *Flow Gated Network*, et ECO.

Les contenus des vidéos ne sont pas les mêmes (mouvement, variations, etc.). Nous allons pour cela adapter la quantification pour chaque vidéo. Le défi majeur de cette adaptation est de trouver un meilleur k_{quanti} pour chaque vidéo sans oublier que le cœur de la méthode se base sur l'égalité des valeurs successives. Pour cela, le k_{quanti} ne doit pas être

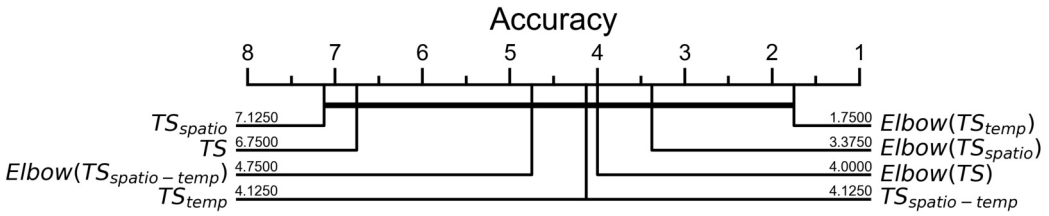
Tableau 4.5 – Taux de classification obtenus sur la classification des vidéos avec la loi du coude [68] (les meilleurs résultats sont en gras).

k_{Quanti}	**Caractéristique**	*RWF2000*	*Movies fights*	*Hockey fights*	*Crowd Violence*
Loi du coude : 4–10	TS	82.9	98.5	**89.3**	82.8
	TS_{temp}	**91.3**	99.0	88.6	**85.3**
	TS_{spatio}	84.8	**100.0**	88.6	82.8
	$TS_{spatio-temp}$	83.9	98.0	88.9	80.4

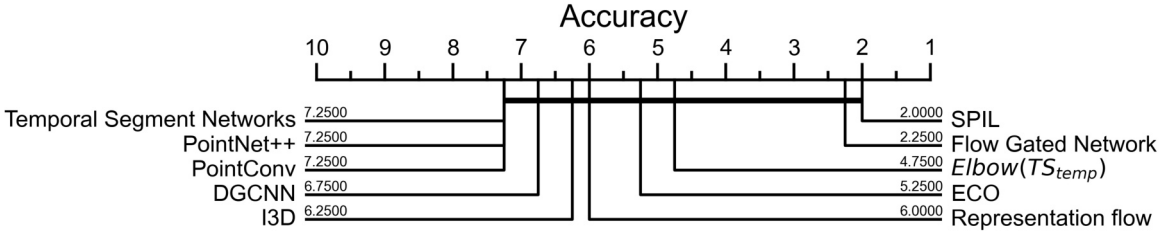
trop grand. Dans l'état-de-l'art du regroupement des données, deux méthodes principales déterminent le bon nombre de groupes, Elbow [68] et silhouette [108]. La méthode Elbow [68] applique l'algorithme k-Moyenne sur les données avec différentes valeurs de k. Ensuite pour chaque k, la distorsion moyenne est calculée entre chaque point et le centre qui lui est attribué. En affichant la courbe des scores en fonction des k, la meilleure valeur de k se trouve au coude de cette courbe (le point où la distorsion moyenne ne diminue pas significativement). La méthode silhouette [108] mesure la qualité du résultat de k-Moyenne. Cette mesure est basée sur la différence entre la distance moyenne de tous les points d'un même groupe avec les distances moyennes des autres groupes. Dans notre cas, nous utilisons la méthode Elbow dans les prochaines expérimentations.

Nous allons maintenant calculer les caractéristiques de stabilité sur les vidéos en se basant sur le choix du k_{quanti} de la méthode Elbow [68] sachant que nous avons choisi une plage de k_{quanti} appartenant à l'intervalle $[4, 10]$. De cette façon, chaque vidéo aura son propre k_{quanti}. Néanmoins, la méthode Elbow n'a pas pu trouver un k_{quanti} optimisé pour certaines vidéos dans les bases *Hockey Fights* et *RWF2000*. Étant donné que la méthode de stabilité est basée sur la capacité de compression de la STI, nous avons fixé k_{Quanti} à 4 pour ces vidéos. Nous utilisons ici le même protocole que dans les expérimentations précédentes. Le tableau 4.5 présente les résultats obtenus. On peut remarquer que la majorité des scores des bases *Movies Fights*, *Crowd Violence* et *RWF2000* ont été améliorés. Mais ceux de *Hockey Fights* se sont dégradés sauf celui de TS. De cette expérimentation, nous remarquons que les hyper-paramètres (k_{quanti} et le choix du CNN ici) influent sur la qualité des résultats.

Malgré les relaxations proposées, il nous est difficile de dire quelle est la caractéristique de stabilité qui permet d'obtenir le meilleur score. Pour cela, nous utilisons diagramme de différence critique [37] afin de pouvoir choisir la meilleure caractéristique de stabilité, c'est-à-dire, comparer les scores obtenus avec k_{quanti} égal à 4 et les scores obtenus avec la loi du coude et ce pour toutes les caractéristiques de stabilité (avec / sans relaxation). La figure 4.15.a présente le diagramme obtenu sur tous les scores de stabilité. Ce dernier illustre que les trois premières méthodes sont celles qui utilisent la loi du coude. Cela montre l'intérêt et l'efficacité de l'utilisation de la loi du coude pour le choix du k_{quanti}. Nous gardons ensuite la meilleure méthode et nous comparons avec les méthodes de l'état-de-l'art. Le nouveau diagramme obtenu est illustré sur la figure 4.15.b. Ce diagramme indique que notre méthode



(a) Comparaison des résultats de notre méthode en utilisant la loi du coude (*Elbow*)



(b) Comparaison des résultats de notre méthode avec les méthodes de l'état-de-l'art

Figure 4.15 – Diagrammes de différence critique obtenus sur l'ensemble des bases pour la classification de vidéos.

basée sur la loi du coude (notée $Elbow(TS_{temp})$) a pu dépasser la méthode ECO et se classe maintenant en troisième position. Néanmoins, les méthodes $SPIL$ et *Flow Gated Network* restent meilleures que la méthode proposée.

4.7 Bilan scientifique

Dans ce chapitre, nous avons présenté une approche qui permet d'extraire des caractéristiques spatio-temporelles en se basant sur une mesure de stabilité. Le calcul de telles caractéristiques est basé sur un algorithme de compression appelé *Run Length Encoding* (RLE), appliqué sur le cube de données d'images, conduisant à une nouvelle représentation intermédiaire de la STI à partir de laquelle les caractéristiques de stabilité peuvent être mesurées. La notion de stabilité a impliqué l'étude de la répétition des valeurs successives. Nous avons étudié également la notion d'égalité et son niveau d'application. Nous avons proposé des approximations du RLE en relâchant les contraintes temporelles ou spatiales dans les prédicats impliqués. À partir de ces approximations du RLE, nous avons ensuite proposé la définition des caractéristiques de stabilité spatio-temporelles.

Les caractéristiques proposées peuvent être utilisées de deux manières. La première consiste à les combiner en une seule image qui résume la STI pour des fins par exemple de visualisation. La deuxième façon est d'utiliser le résumé pour alimenter un classificateur et de procéder, par exemple, à une classification. Ces deux types d'utilisation ont été expérimentés dans deux problèmes différents : l'analyse des $STIS$ et l'analyse de vidéos. Nous avons constaté que le résumé des $STIS$ a facilité l'interprétation des territoires urbains

détectés. Par contre pour les vidéos, tout dépend de la vitesse de déplacement des objets ou de la caméra dans la scène. Les résultats visuels entre les bases *Movies Fights* et *Crowd Violence* montrent les limites du résumé. Les caractéristiques ont aussi été impliquées dans un problème de classification binaire. La première expérimentation concerne l'analyse de la couverture du sol avec les $STIS$. Les résultats obtenus avec un simple arbre de décision avec les caractéristiques de stabilité sont remarquables et atteignent une précision très proche de $TempCNN$. Par le simple fait que nous combinons les caractéristiques de stabilité avec les pixels temporels et que nous utilisons une forêt aléatoire, nous atteignons un plateau de scores. En revanche dans la deuxième expérimentation qui consiste à classifier les vidéos, nous obtenons des scores encourageants mais qui ne sont pas toujours en première position.

L'avènement des méthodes basées sur les réseaux de neurones profonds a permis de fusionner les deux étapes : extraction de caractéristiques et décision automatique. Un tel système effectue simultanément l'extraction de caractéristiques et la prise de décision automatique sur plusieurs couches. La suite de nos recherches se focalise alors sur l'utilisation des méthodes basées sur les réseaux de neurones profonds, en particulier les réseaux de neurones convolutifs, pour d'un côté extraire des caractéristiques tout en classifiant les données. Cela va permettre d'optimiser la qualité des caractéristiques en fonction du problème ciblé et de s'affranchir des caractéristiques *hand-crafted* qui requiert un expert du domaine pour pouvoir les définir.

Les travaux présentés dans ce chapitre ont fait l'objet de deux publications dans des conférences internationales :

- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Spatio-temporal stability analysis in Satellite Image Times Series.** *Second International Conference on Pattern Recognition and Artificial Intelligence.* 2020, pages 484–499.
- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Urban land cover analysis from satellite image time series based on temporal stability.** *Joint Urban Remote Sensing Event.* 2019, pages 1–4.

Chapitre 5

Étude des variations des séquences temporelles d'images

Pour examiner la vérité, il est besoin, une fois dans sa vie, de mettre toutes choses en doute autant qu'il se peut.

– René Descartes

Ce chapitre présente une méthode dont l'objectif est d'analyser les variations des STI à partir de représentations planaires. La section 5.1 introduit ce chapitre sur l'utilisation des méthodes d'apprentissage automatiques basées sur les CNN. La méthode proposée,

nommée *Deep Spatio-Temporal Representation* ($Deep-STaR$) est détaillée dans la section 5.2. $Deep-STaR$ a pour but de transformer une STI en plusieurs représentations $2D$ incluant l'information spatiale et temporelle simultanément. Les sections 5.3 et 5.4 exposent respectivement l'apprentissage des caractéristiques spatio-temporelles avec un CNN $2D$ et la technique utilisée pour la prise de décision. L'interprétabilité des CNN est une parmi les problématiques clé du domaine de l'apprentissage automatique. Dans ce contexte, la section 5.5 présente deux approches permettant d'interpréter les CNN. La première approche se base sur les cartes de saillance afin d'analyser les STR et propose deux mécanismes d'attention. Un qui analyse le domaine temporel le plus utilisé par le CNN et l'autre qui analyse le domaine spatial des STR et génère une carte de segmentation sémantique associée au domaine spatial de la STI. La deuxième approche permet d'analyser la nature des filtres appris après la fin de l'entraînement du CNN. Les études expérimentales et le bilan scientifique sont respectivement présentés dans les sections 5.6 et 5.7.

5.1 Introduction

Dans le chapitre précédent, nous avons défini et impliqué des caractéristiques de stabilité dans deux problèmes de classification qui sont l'analyse de la tache urbaine et la détection de la violence à partir de vidéos. Cependant, les caractéristiques de la stabilité ont certaines limites et ne sont pas discriminantes dans le cadre d'objets en mouvement dans la scène que nous pouvons voir dans les différents résumés. Une telle déformation est aussi due à la quantification des valeurs de la STI. L'enjeu est de maintenant analyser non seulement la stabilité sur les vraies valeurs de la STI (non quantifiées) mais aussi le mouvement des objets dans la scène conduisant à une analyse des variations sur les dimensions temporelle et spatiale de la STI. Différentes études ont démontré que les variations peuvent être utilisées dans diverses applications. Par exemple, dans le cadre de la suppression de la distorsion spatiale afin d'améliorer la qualité des vidéos [92] ou le suivi du mouvement de foules [40].

Dans ce chapitre, nous présentons la méthode proposée, nommée Deep-STaR comme *Deep Spatio-Temporal Representation*. Cette dernière utilise les méthodes d'apprentissage automatiques basées sur les réseaux de neurones profonds pour effectuer une telle analyse afin que les caractéristiques extraites soient optimisées pour un problème particulier. Nous nous intéressons précisément aux réseaux de neurones convolutionnels (CNN) où les convolutions peuvent s'appliquer de plusieurs manières et ce selon le type de données. Dans le cas de la classification des images [31, 73], ce sont des convolutions $2D$ qui sont utilisées. Quand des séries temporelles de pixels sont considérées [98], alors ce sont des convolutions $1D$. Enfin, des convolutions $3D$ sont naturellement utilisées dans le cas des STI [64]. Pour résumer les méthodes de l'état-de-l'art, la nature des caractéristiques extraites dépend des domaines dans lesquels les convolutions sont appliquées. Cependant, certaines méthodes de l'état-de-l'art produisent des informations spatio-temporelles mais elles souffrent de certaines limites. D'un coté, les CNN $3D$ traitent les données de façon locale (*e.g.*, 16 images [64]) et ils requièrent de grandes masses de données afin de fixer tous

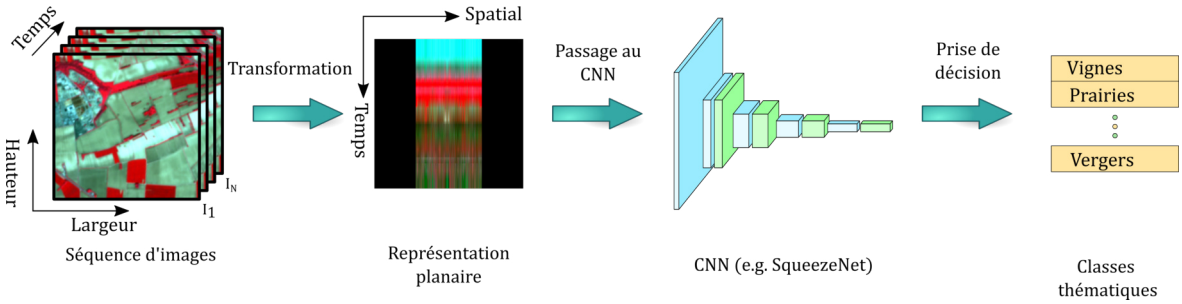


Figure 5.1 – La chaîne de traitement globale de $Deep - STaR$.

les paramètres du modèle. D'autre part, les méthodes basées sur les nuages de points nécessitent un calcul en amont des points d'intérêt, par exemple via une détection des personnes [128], ce qui peut être coûteux d'un point de vue algorithmique et entraîner des erreurs.

Nous proposons dans ce chapitre une méthode d'extraction des caractéristiques spatio-temporelles avec seulement des modèles dotés de convolutions $2D$. Pour ce faire, nous proposons de transformer les données de STI, initialement définies dans un espace $2D+t$, vers des représentations planaires $2D$ qui capturent des informations spatiales et temporelles. De telles représentations permettent à un CNN $2D$, d'apprendre des poids de filtres $2D$ permettant d'extraire des caractéristiques qui sont nativement spatio-temporelles. Nous bénéficierons aussi, grâce à de telles représentations, des connaissances des modèles déjà entraînés sur de grandes bases telles que ImageNet [31]. La figure 5.1 présente la chaîne de traitement globale de la méthode. En plus de cette représentation originale, nous proposons un nouveau mécanisme d'attention *post-hoc* qui permet l'explication des différentes décisions prises par le CNN et aussi la création de cartes sémantiques du domaine spatial original de la STI. Sa particularité est d'intégrer les informations d'attention dans l'espace original de la STI afin de mieux comprendre la décision prise par le CNN.

La suite de ce chapitre va être organisée comme suit. La section 5.2 présente la méthode proposée concernant la construction des représentations planaires et les stratégies de conservation de l'information spatiale. La section 5.3 introduira les CNN utilisés pour apprendre les caractéristiques. Le processus de décision sera présenté dans la section 5.4. La section 5.5 expliquera quelques méthodes permettant la compréhension de la décision prise par le CNN. Enfin, l'étude expérimentale et le bilan scientifique seront présentés respectivement dans les sections 5.6 et 5.7.

5.2 Méthode proposée : $Deep - STaR$

Généralement, la classification des STI peut se faire avec deux méthodes différentes. La première se base sur l'utilisation des CNN $3D$ et la deuxième considère les pixels comme des séries temporelles indépendantes. La méthode $Deep-STaR$ proposée dans ce chapitre consiste à analyser la STI d'un point de vue intermédiaire, c'est-à-dire entre les données $2D+t$ et les séries temporelles (pixels temporels) qui sont des données unidimensionnelles.

Pour ce faire, nous enrichissons les pixels temporels avec de l'information spatiale. La réalisation d'une telle tâche repose sur la construction d'une représentation planaire contenant à la fois les informations spatiales et temporelles. Le cœur principal d'une telle représentation se localise dans la transformation du domaine spatial $\mathcal{D}$ défini dans un espace $2D$ vers un espace $1D$, tout en préservant une configuration spatiale significative, c'est-à-dire en conservant partiellement le voisinage des pixels. Enfin, la combinaison de la dimension spatiale réduite avec la dimension temporelle conduit à la création d'une représentation spatio-temporelle $2D$ d'un pixel temporel, noté STR (*Spatio-temporal Representation*).

Une fois que cette stratégie de création des STR est définie, la suite consiste à représenter une STI par un ensemble de STR et à entraîner un modèle de classification de bout-en-bout permettant à la fois d'apprendre des caractéristiques et de conduire à une décision de classification. Avec cette réduction de complexité de représentation, la manipulation et l'interprétation deviennent plus aisées. De plus, les CNN $2D$ vont pouvoir extraire directement des caractéristiques spatio-temporelles avec seulement des convolutions $2D$. La Figure 5.1 présente la chaîne de traitement globale de $Deep - STaR$. Dans la suite, nous décrivons en détail la création des STR et les différentes stratégies élaborées pour la conservation de l'information spatiale.

5.2.1 Représentation des données

Nous expliquons maintenant comment une STR est construite. Avant de transformer le cube $2D + t$, nous commençons par définir comment transformer seulement le domaine spatial $\mathcal{D}$ défini dans un espace $2D$ vers un espace $1D$. Pour commencer, un pixel est caractérisé par ses valeurs radiométriques et sa position (x, y) dans le domaine spatial $\mathcal{D}$ de l'image. En général, deux topologies de voisinages d'un point sont possibles [41]. Un pixel peut avoir 4 ou 8 plus proches voisins directement connectés, sauf dans le cas où le pixel se situe aux bords de l'image. Dans notre cas, nous utilisons la 8-connexité afin de maximiser l'information spatiale. Afin de transformer le domaine spatial $2D$, nous utilisons des structures $1D$ de longueur L qui limiteront, pour chaque pixel, à deux le nombre de voisins directement connectés à l'exception des pixels qui sont aux extrémités de la structure $1D$. Il faut aussi que ces structures assurent une exploration isotropique de l'espace afin que l'information spatiale conservée soit statistiquement significative.

Les structures $1D$ sont analytiquement définies par une paramétrisation des pixels en fonction de leurs positions relatives dans le domaine spatial $\mathcal{D}$. Ces dernières sont aussi interprétées comme des courbes. Ainsi, l'information spatiale est diminuée puisqu'un pixel dans une STR n'aura que 2 voisins parmi les 8 possibles. Pour cela, nous considérons la courbe Γ définie dans $\mathcal{D}$ et qui sélectionne pour un pixel, d'autres qui lui sont juxtaposés. Le nombre de pixels est limité à la longueur L de Γ. L'origine de Γ est le pixel cible à enrichir avec l'information spatiale. Ensuite, les pixels sont indexés par rapport à l'abscisse curviligne des points sur Γ et le point initial est une des deux extrémités de Γ. Enfin, une séquence $1D$ ordonnée selon les pixels sélectionnés par Γ, notée $(p_1, \ldots, p_L)$ est construite. La figure 5.2 illustre cette construction.

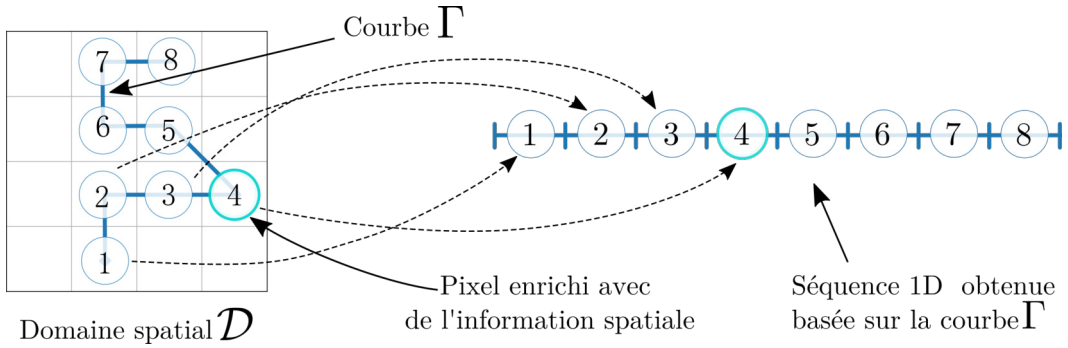


Figure 5.2 – Transformation partielle du domaine spatial $\mathcal{D}$ en un vecteur $1D$.

Une fois la transformation définie, la considération de la STI change de point de vue. Au lieu de regarder les images dans leur vrai domaine spatial $\mathcal{D}$ ou au lieu de regarder les pixels comme des séries temporelles, nous nous concentrons sur les entités sur lesquelles la représentation de la dimension spatiale $\mathcal{D}$ est réduite. La figure 5.3 illustre une STI, un pixel temporel extrait de la STI et le point de vue adopté par la méthode proposée pour considérer des données $2D + t$.

La construction de la STR est ensuite réalisée en appliquant le processus de réduction du domaine spatial $\mathcal{D}$ à chacune des T images de la STI avec la même courbe Γ. De cette façon, nous obtenons T vecteurs $1D$, notés $p = \{(p_1^i, \ldots, p_L^i)\}_{i=1}^T$, qui sont par la suite empilés verticalement afin d'être les lignes d'une matrice $2D$. Cette matrice est interprétée comme une image et représente la STR. Le nombre de lignes de la STR est le nombre d'images T de la STI ou la longueur des pixels temporels $((p_t)_{t=1}^T)$. La largeur de la STR est la longueur L de la courbe Γ. Notons que si les pixels de la STI originale étaient caractérisés par des valeurs radiométriques multiples (*i.e.*, images multi-bandes), alors la STR est une image $2D$ multi-bandes. La figure 5.4 illustre un exemple de construction d'une STR avec la courbe Γ.

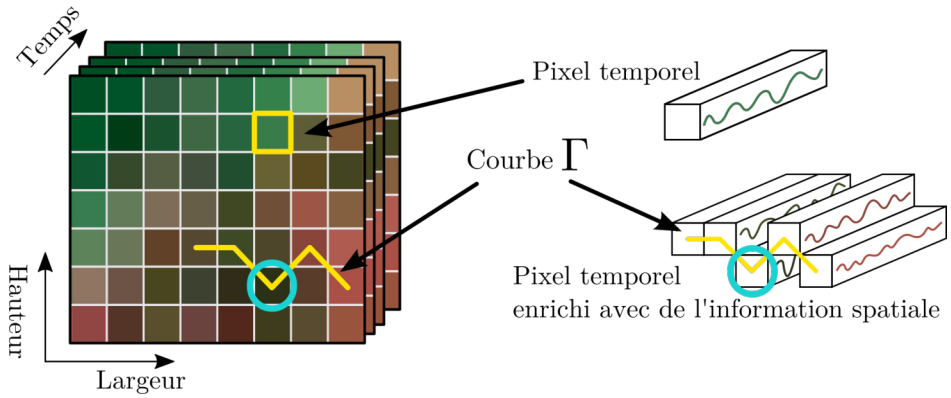


Figure 5.3 – Illustration de trois techniques de gestion des données ; (à gauche) représentation originale $2D + t$ comme un cube ; (à droite - haut) pixel temporel ; (à droite - bas) pixel temporel encerclé enrichi avec l'information spatiale selon la courbe Γ.

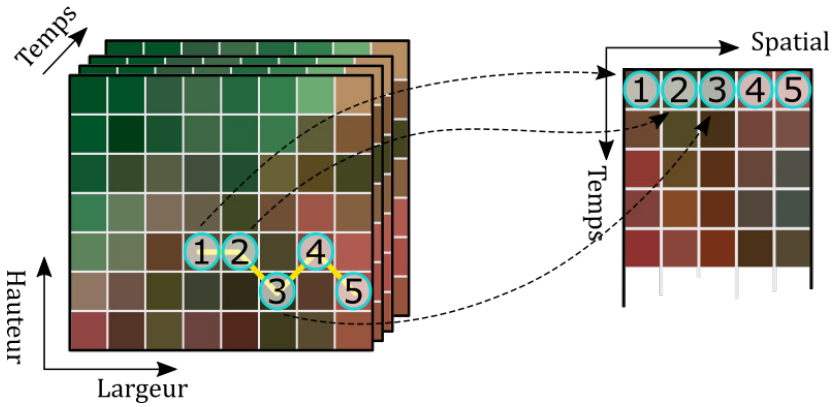


Figure 5.4 – Illustration d'un exemple de construction d'une STR avec la courbe Γ.

Les courbes permettant une telle conservation de l'information spatiale sont diverses. Dans la suite, nous présenterons différentes stratégies pour tracer des courbes dans le domaine spatial $\mathcal{D}$ de l'image. Ces courbes peuvent conserver l'information via deux stratégies différentes, soit de façon globale ou soit de façon locale.

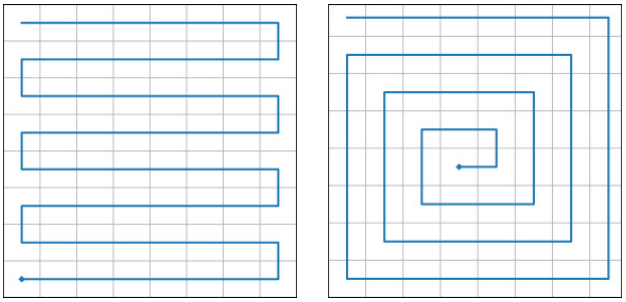
5.2.2 Stratégies de conservation de l'information spatiale

Dans cette section, nous présentons les deux stratégies de conservation de l'information spatiale que nous utilisons pour la construction des STR. La première stratégie est basée sur l'utilisation d'une courbe couvrant le domaine spatial $\mathcal{D}$ de façon globale, notée $G-STR$ pour Globale STR. La deuxième stratégie est basée sur des courbes locales qui permettront de représenter la STI initiale avec plusieurs STR, notée $MS-STR$ pour Multi-Segments STR. Ces deux stratégies sont respectivement présentées dans les sections 5.2.2.1 et 5.2.2.2.

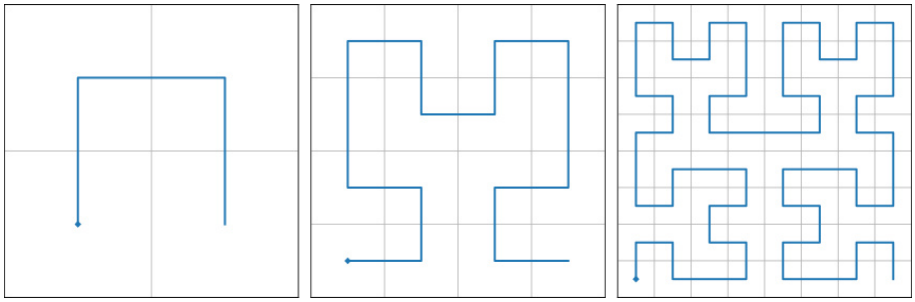
5.2.2.1 Courbes globales remplissant l'espace

Le premier type de courbes permet de transformer le domaine spatial $\mathcal{D}$ de façon globale conduisant à une représentation spatio-temporelle globale ($G-STR$). En mathématiques, une courbe de remplissage est un chemin qui passe par tous les points d'un espace de façon continue sans jamais se croiser. Ce type de courbes peut être vu comme un moyen d'ordonnancement des pixels afin de réduire le domaine spatial à une seule dimension [17]. Ces courbes partagent une propriété importante qui est la préservation de la localité des pixels dans le nouvel espace transformé par rapport au domaine original. Grâce à cette propriété, la reprojection des pixels dans le domaine initial peut être faite avec précision. Ces courbes conservent l'information spatiale la plus représentative au moment du passage d'un espace $2D$ à $1D$.

En 1890, un mathématicien a proposé la première courbe de remplissage et elle fut nommée, en son nom, courbe de Peano. Un an plus tard, la courbe de Hilbert est proposée.



(a) $\Re_{snake}$ (b) $\Re_{spiral}$



(c) $\Re_{Hilbert}$ (trois premier ordres)

FIGURE 5.5 – Courbes de remplissage utilisées pour transformer une image $2D$ vers un vecteur $1D$ de pixels.

Elle se construit de manière alternative à la courbe précédente. Ces courbes sont des cas particuliers des courbes basées sur les fractales. En outre, il existe d'autres courbes remplissant l'espace telles que la courbe serpent ou spirale. Chacune de ces courbes a sa propre stratégie pour garder des voisins statistiquement représentatifs sans biais. Dans notre cas, nous avons sélectionné trois courbes différentes qui sont celles de serpent, de spirale et de HILBERT.

Avant de décrire les courbes sélectionnées, nous rappelons que les pixels sont initialement définis par leur position (x, y) dans $\mathcal{D}$. En appliquant la transformation avec l'une de ces courbes, les pixels seront donc représentés par seulement un indice j entier. Pour cela, nous définissons la fonction $\Re$ de la transformation considérée :

$$\begin{array}{rccc} \Re : & \mathcal{D} & \rightarrow & [0, \mathbb{W} \times \mathbb{H} - 1] \\ & (x, y) & \mapsto & j = \Re(x, y) \end{array} \tag{5.1}$$

qui associe à un pixel (x, y) un indice j dans la courbe Γ.

Courbe serpent

La courbe serpent, en anglais *Snake* et notée $\Re_{snake}$, est la plus naïve des courbes. Le remplissage de l'espace $\mathcal{D}$ avec $\Re_{snake}$ est réalisé en parcourant l'espace consécutivement ligne par ligne comme un serpent. Afin de toujours avoir des pixels voisins directement

connectés, les liens entre les lignes sont faites de façon intelligente. Les bouts des lignes impaires sont liés aux têtes des lignes paires, et *vice versa*. La figure 5.5.a illustre la courbe $\Re_{snake}$.

Courbe spirale

L'idée principale de cette courbe est basée sur la spirale d'Archimède, notée $\Re_{spiral}$. La courbe spirale $\Re_{spiral}$ remplit un espace carré en commençant au centre lié à son voisin de droite puis elle tourne autour du centre tout en s'éloignant. Afin de construire cette courbe, nous introduisons les variables dx, dy qui permettent d'indiquer le prochain point par rapport à la position actuelle. dx, dy sont initialisées à 0 et 1. Le prochain point est obtenu avec une simple addition des coordonnées avec les deux variables, comme suit $(x+dx, y+dy)$. Les points angulaires sont ceux vérifiant $(x = y)$, $(x = -y$ et $y > 0)$, $(x-1 = -y$ et $x > 0)$ ou $(y = 1 - x$ et $y > 0)$. La courbe doit aller à droite, à gauche, en bas ou en haut selon les directions de (dx, dy) qui sont respectivement $(0, 1), (-1, 0), (0, -1)$ et $(1, 0)$. La figure 5.5.b illustre la courbe $\Re_{spiral}$.

Courbe de Hilbert

La troisième courbe est celle de Hilbert, notée $\Re_{Hilbert}$. Cette dernière est parmi les courbes remplissant l'espace qui sont auto-similaires [17]. La construction de $\Re_{Hilbert}$ se fait de manière récursive en divisant d'abord le domaine carré en 4 parties carrées égales. Puis ces quatre parties sont liées en vérifiant la condition suivante : « *deux parties avec une arête commune ont deux indexes consécutifs* ». Ce processus (division et lien) est appliqué de manière récursive sur les carrés dont la largeur est une puissance de 2. $\Re_{Hilbert}$ a été utilisée dans le contexte d'indexation pour la recherche d'images [90]. La figure 5.5.c présente les trois premiers ordres de $\Re_{Hilbert}$.

5.2.2.2 Courbes locales

Le but principal de cette stratégie est d'analyser les données d'un point de vue intermédiaire entre le niveau du pixel temporel et celui de la STI. L'idée de cette stratégie est de générer plusieurs segments $1D$ (courbes) dans le même domaine spatial $\mathcal{D}$. Ces segments vont conserver une information spatiale locale qui est limitée à L voisins du pixel initial. Ensuite, pour une STI donnée, différentes STR sont construites à partir des segments générés.

Une telle stratégie a un avantage très important pour les méthodes d'apprentissage supervisées qui nécessitent souvent un grand ensemble de données pour leur entraînement. D'un point de vue local, un pixel temporel est enrichi avec l'information spatiale en se basant sur la courbe générée. Mais en générant plusieurs courbes passant par le même pixel, plusieurs STR caractérisant différentes informations spatiales sont créées. Ainsi, la base

Figure 5.6 – Illustration de différents segments de RW. Les points sur les extrémités des courbes sont leurs points initiaux.

des STR devient plus grande. La stratégie élaborée pour générer les segments est expliquée ci-après.

Segments basés sur les marches aléatoires

La marche aléatoire [45], dite en anglais *Random Walk* et notée RW, est un processus mathématique basé sur un système itératif aléatoire. Les itérations suivent toutes les propriétés Markoviennes d'ordre 1. Dans notre cas, les RW sont utilisées pour générer une courbe aléatoire dans le plan $\mathcal{D}$ de l'image, une telle courbe ayant une longueur L est notée $RW(L)$. Le premier point de la courbe est choisi aléatoirement sur $\mathcal{D}$ et pour le point suivant, 8 directions sont possibles sauf si un pixel est proche de la frontière de la région. Grâce au processus aléatoire, pour une STI donnée, N_{seg} différents STR sont générées avec des initialisations différentes des courbes. Elles modélisent aussi la STI avec plusieurs représentations spatio-temporelles multi-segments ($MS-STR$). N_{seg} est un paramètre permettant de contrôler le nombre de représentations. La figure 5.6 illustre différents segments de RW.

Jusqu'à présent, les STR peuvent être générées de deux manières. De façon locale ($MS-STR$) et globale ($G-STR$). Maintenant, nous allons passer à l'étape d'apprentissage des caractéristiques à l'aide des modèles des réseaux de neurones profonds.

5.3 Apprentissage automatique des caractéristiques spatio-temporelles via des convolutions

Les méthodes d'apprentissage profond sont adaptées à apprendre comment prendre une décision de façon automatique. Dans notre cas, nous allons utiliser les réseaux de neurones convolutifs (CNN) pour apprendre les caractéristiques à partir des images de STR données en entrée. Les STR contiennent à la fois des informations spatiales et temporelles. Cela permet aux CNN d'apprendre directement des caractéristiques spatio-temporelles

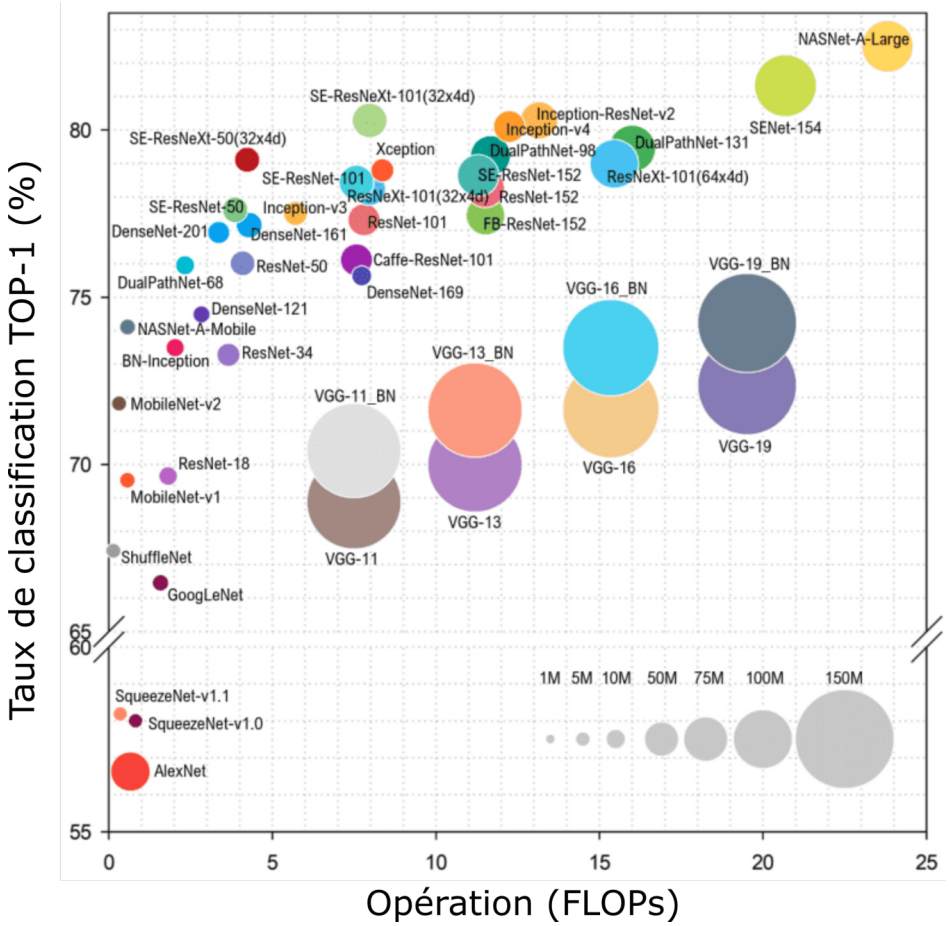
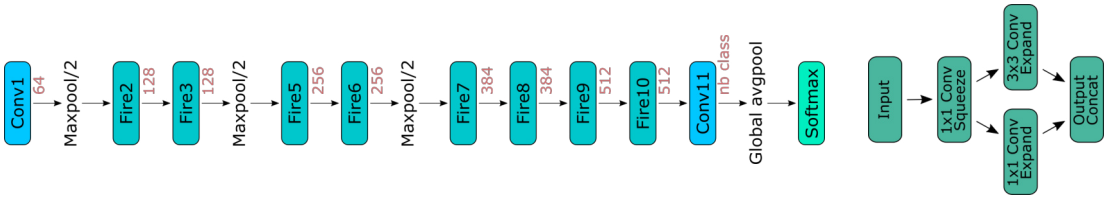


Figure 5.7 – Comparaison de différents CNN selon leur nombre d'opérations flottantes par seconde (FLOPs) et leur taux de classification (TC) du TOP-1 lors du test sur ImageNet. L'air du disque indique le nombre de paramètres du modèle. Figure prise de [13].

adaptées à une problématique grâce à la fonction d'optimisation (*e.g.*, entropie ou l'erreur quadratique moyenne). Quant aux couches d'activation (*e.g.*, sigmoïde, ReLU), de *pooling* et de normalisation du lot qui viennent après les couches de convolutions, elles sont conçues pour ne sélectionner que les caractéristiques d'ordre supérieur à partir de l'entrée tout en réduisant progressivement la taille spatiale des entrées et le nombre de paramètres à calculer dans le réseau. Enfin, la décision est prise à l'aide d'une couche entièrement connectée qui a le même principe qu'un perceptron multicouche. Le vecteur de sortie est procédé avec la fonction *softmax* qui permet de donner un vecteur que l'on peut interpréter comme une probabilité, par lequel l'étiquette de classe des données d'entrée est prédite.

Le principe de notre méthode est de transformer la représentation des données afin de pouvoir utiliser les CNN initialement conçus pour traiter les images $2D$. Cependant, une grande variété de CNN existe avec des architectures qui diffèrent d'un modèle à l'autre. La figure 5.7 présente une comparaison d'un ensemble de modèles avec leurs nombres d'opérations flottantes par seconde (FLOPs) et le taux de classification TOP-1 lors de la classifi-



(a) (b)

Figure 5.8 – L'architecture du CNN SqueezeNet V1.1 ; (a) le modèle complet ; (b) la couche Fire.

cation des images de la base ImageNet [31, 73, 13]. Nous remarquons que AlexNet est le modèle qui donne les résultats les plus faibles avec un grand nombre de paramètres. VGG16 explose en nombre de paramètres mais atteint un taux de reconnaissance remarquable. Par contre, VGG16 a été surpassé par les modèles ResNet-50, Densenet-161 et Inception V3 avec des architectures plus légères. Les modèles les plus légers sont SqueezeNet V1.1 [58], ShuffleNet V2 x1.0, GoogleNet et MobileNet V2 qui donnent des scores qui se situent entre ceux de AlexNet en dernière position et MobileNet V2 en première position.

Dans notre cas, nous allons utiliser le modèle SqueezeNet V1.1 [58] qui est présenté dans la figure 5.8. Ce modèle est au même niveau de classification que AlexNet quand il est évalué sur ImageNet [31, 73] avec 50 fois moins de paramètres. SqueezeNet propose une couche nommée Fire qui applique une convolution 1×1 suivie de deux convolutions différentes, une avec un filtre 1×1 et l'autre un filtre 3×3. Enfin les résultats de ces deux convolutions sont concaténés. La partie du modèle qui extrait des caractéristiques va de la couche « Conv1 » jusqu'à la couche « Fire10 », comme illustré dans la figure 5.8. La taille des caractéristiques apprises dans l'espace latent est de 13×13 projetée sur 512 dimensions. Le classificateur est composé de trois couches qui sont : une convolution qui a pour but de réduire la dimension de l'espace latent au nombre de classes, une couche qui fait une moyenne globale pour chaque dimension afin d'avoir une valeur pour chacune des classes et enfin la *softmax* est appliquée pour transformer ces valeurs en des probabilités dont la somme est égale à 1.

Le modèle choisi sera entraîné avec les différentes STR (locales ou globales) que nous avons proposées. Étant donné que les STR sont des images $2D$, le modèle peut alors bénéficier des poids déjà appris à partir d'autres bases telles que ImageNet [31, 73] puis ces poids sont raffinés en entraînant le CNN sur nos images.

5.4 Prise de décision

Dans le contexte de la classification d'images, le modèle retourne les « probabilités » Y^c associées à chacune des classes, avec c inclus dans l'intervalle $[1, C]$ et C est le nombre de classes. Ensuite, le label où la probabilité est maximum est affecté à l'image. Dans le cas des STI, la décision peut être prise de différentes manières qui dépendent de la stratégie utili-

sée. Pour les $CNN\ 3D$, une STI est d'abord coupée en plusieurs STI de 16 images. Pour les $CNN\ 1D$ ou temporels, la STI est considérée comme un ensemble de pixels temporels. À ce moment l'entrée est soit un pixel temporel ou une STI de 16 images, selon le modèle considéré. La décision pour ces deux types de CNN peut être prise de deux manières distinctes. Le premier type de décision procède dans l'espace latent des caractéristiques, c'est-à-dire, les caractéristiques extraites sont agrégées progressivement (*i.e.*, addition, fusion ou multiplication point à point), noté $AggC$, puis le résultat est donné au classificateur du modèle. Le deuxième type de décision consiste à prendre une décision, dite locale, pour chaque entrée. Puis, la décision globale de la STI est faite en calculant la moyenne des probabilités retournées pour chaque classe de toutes les entrées du modèle. Dans les deux stratégies de décision, le label affecté à la STI est celui où la probabilité est maximum.

La méthode proposée permet de représenter une STI par une ou plusieurs STR selon la stratégie utilisée, $G-STR$ ou $MS-STR$. Quand l'approche locale $MS-STR$ est considérée, de nombreuses courbes peuvent être extraites du domaine $\mathcal{D}$ de la STI conduisant à N_{seg} STR. Par contre dans l'approche globale $G-STR$, une STI est représentée avec des STR spécifiques. La stratégie de décision peut être une des deux présentées précédemment ($AggC$ ou locale). Dans notre cas, nous utilisons la stratégie de décision locale. Le CNN fournit une décision pour chaque STR, en donnant les probabilités des classes. Ensuite, la décision globale de la STI est prise en faisant la moyenne des probabilités retournées pour chaque classe puis nous affectons le label ayant la probabilité la plus haute. La figure 5.9 présente le processus de prise de décision.

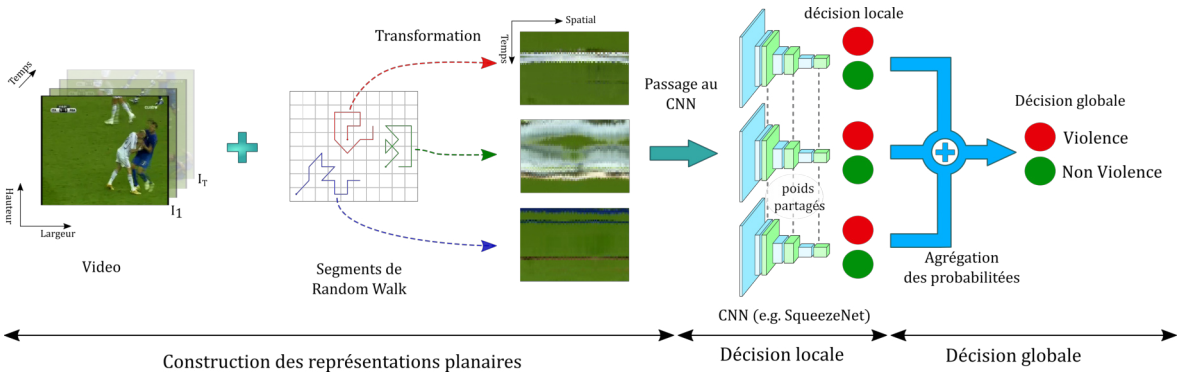


Figure 5.9 – Illustration du processus global de la prise de décision avec la stratégie locale $MS-STR$.

Après la prise de décision, nous nous intéressons à l'explicabilité des CNN qui ont permis la prise de décision, en fonction des caractéristiques spatio-temporelles extraites. Une telle analyse permet aussi de capter, en fonction des caractéristiques et du classificateur, les régions dans l'image qui ont été les plus saillantes pour la décision.

5.5 Explication des décisions prises par les CNN

Les méthodes basées sur les réseaux de neurones profonds, plus précisément les CNN, sont souvent considérées comme des boîtes noires en raison de leur manque d'interprétabilité. De tels modèles sont entraînés en fixant des millions de paramètres afin de s'intéresser aux régions des images (ou des signaux) les plus discriminantes. L'explication de la décision des CNN reste complexe à cause des millions de paramètres qui les constituent. La figure 5.7 présente le nombre de paramètres de plusieurs CNN $2D$. Le modèle $2D$ le plus petit a déjà un million de paramètres. Les modèles $3D$ quant à eux explosent en nombre de paramètres. Ce qui va augmenter la difficulté d'explication de leurs décisions. Dans ce contexte, certains chercheurs ont proposé des méthodes d'attention permettant de générer des cartes de saillance par rapport à la décision du modèle. Ces cartes mettent en évidence les régions de l'image les plus pertinentes qui ont contribué à la prise de décision. Ces méthodes sont utilisées en *post-hoc*, c'est-à-dire après l'entraînement, à opposer aux méthodes qui intègrent une couche d'attention où ces paramètres sont fixés au cours de l'entraînement [79].

La première méthode produisant des cartes de saillance a été proposée en 2016, elle a introduit la notion de *Class Activation Map* (CAM) [151] pour générer une carte de saillance S^c pour la classe, sachant que c est inclus dans l'intervalle $[1, C]$ avec C le nombre de classes. L'approche des CAM pondère les caractéristiques extraites A de dimensions $h \times w \times K$ dans l'espace latent avec les poids w de la couche de convolution qui est avant celle de la moyenne globale, *Global Average Pooling* (GAP). La hauteur h et la largeur w de A sont strictement inférieures à celles de l'image d'entrée et K est le nombre de filtres. De cette façon, les CAM identifient l'importance de chaque dimension des caractéristiques. L'équation 5.2 permet le calcul des CAM :

$$S^c(x, y) = \sum_{k=1}^{K} w_k^c A^k(x, y), \quad \forall x \in [1, h], \forall y \in [1, w] \tag{5.2}$$

Afin de superposer S^c sur l'image, il suffit d'augmenter les dimensions de S^c à la taille originale de l'image. Les CAM sont également utilisées lors de l'analyse des CNN avec des convolutions temporelles [37]. Cependant, le calcul des CAM se limite seulement aux modèles qui sont dotés d'une couche GAP, avant la couche du classificateur et du *softmax* [151]. Les CAM sont optimisées en utilisant une rétro-propagation guidée des poids w par rapport à la probabilité Y^c trouvée par le CNN. Cette optimisation rend compatible le calcul des cartes de saillances S^c avec les CNN se terminant par une couche entièrement connectée. Deux méthodes optimisant les CAM sont proposées. La première méthode est $GradCAM$ [118]. Cette dernière rétro-propage le gradient et le normalise en le divisant par le nombre de caractéristiques $h \times w$ de A^k. Le calcul du $GradCAM$ se fait avec l'équation 5.2 mais avec des poids raffinés. L'équation 5.3 définie ci-dessous montre comment calculer les nouveaux poids :

$$w_k^c = \frac{1}{h \times w} \sum_{i}^{h} \sum_{j}^{w} \frac{\partial Y^c}{\partial A_{ij}^k} \tag{5.3}$$

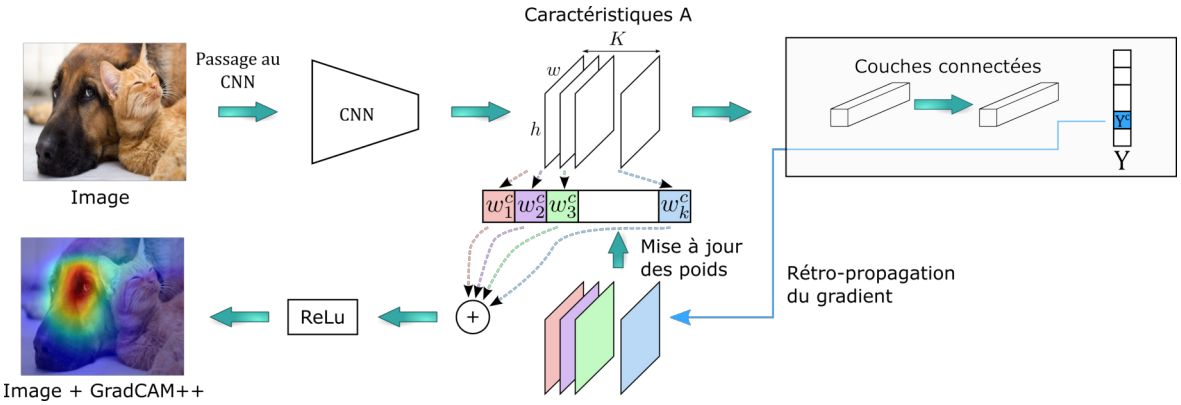


Figure 5.10 – Illustration de la méthode du calcul du $GradCAM++$ [21].

La deuxième méthode est $GradCAM++$ [21]. Cette dernière applique la fonction d'activation *ReLu* sur le gradient afin de ne garder que les valeurs positives pour récupérer uniquement l'information la plus saillante. La figure 5.10 présente un schéma du calcul du $GradCAM++$. L'équation 5.4 présentée ci-dessous expose comment calculer les nouveaux poids :

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} ReLu\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \quad \text{avec} \quad \alpha_{ij}^{kc} = \begin{cases} \frac{1}{\sum_a \sum_b \frac{\partial Y^c}{\partial A_{ab}^k}} & \text{si } \frac{\partial Y^c}{\partial A_{ij}^k} = 1 \\ 0 & \text{sinon} \end{cases} \tag{5.4}$$

Récemment, une autre méthode conçue pour améliorer la qualité du GradCAM++ nommée $ScoreCAM$ est présentée dans [137]. Cette dernière optimise le calcul du α_{ij}^{kc} en introduisant la fonction de CIC pour *Channel-wise Increase of Confidence*. CIC n'utilise que les poids de la dernière couche mais elle calcule un score de confiance pour chaque dimension k des caractéristiques latentes A. Pour ce faire, l'image (I) d'entrée est multipliée au niveau pixel avec chaque A^k définissant $CIC(A^k)$ comme suit :

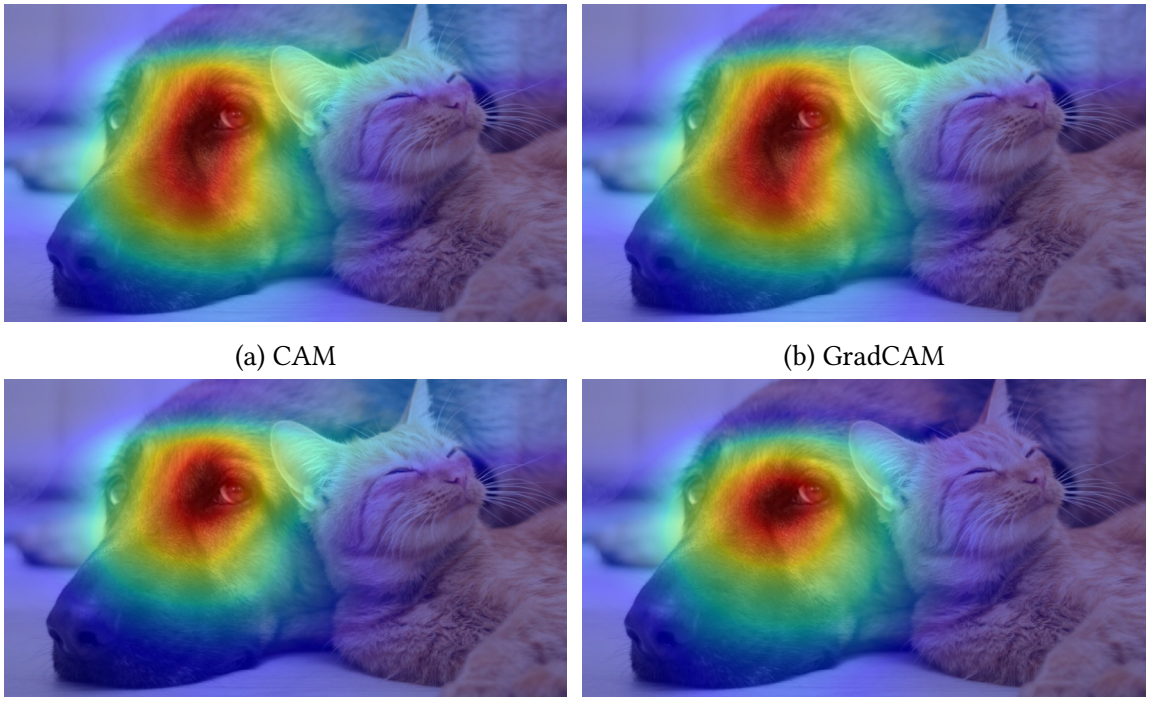
$$CIC(A^k) = f\left(I * Normaliser(Up(A^k))\right) \tag{5.5}$$

où f est le CNN, I est l'image initiale, $Normaliser$ est une fonction de normalisation qui étire les valeurs de l'image entre 0 et 1 selon son minimum et maximum et Up permet d'agrandir la taille de A^k à la même taille que l'image I. Ensuite, $ScoreCAM$ optimise l'équation 5.2 et l'équation 5.6 présente la nouvelle formule pour calculer la carte de saillance :

$$S^c(x,y) = ReLu\left(\sum_k \alpha^{kc} A^k\right)$$
$$\text{avec} \quad \alpha^{kc} = softmax\left(\left[CIC(A^k)\right]_{k=1}^K\right), \forall x \in [1,h], \forall y \in [1,w] \tag{5.6}$$

La figure 5.11 présente des cartes de saillance obtenues avec les différentes méthodes présentées précédemment dans un problème de classification d'images. Nous remarquons

que les cartes de saillance montrent que le modèle se focalise sur la tête du chien justifiant que la probabilité la plus forte est associée à la classe « chien ». En revanche, nous observons qu'il n'y a pas de différence visuelle entre la CAM et la $GradCAM$. Par ailleurs, $GradCAM++$ et $ScoreCAM$ présentent une zone d'attention plus restreinte contrairement aux méthodes précédentes. Nous notons aussi qu'il y a un léger décalage sur la zone d'attention entre $GradCAM++$ et $ScoreCAM$.



(a) CAM (b) GradCAM

(c) GradCAM++ (d) ScoreCAM

Figure 5.11 – Visualisation des différentes cartes de sailliance obtenues avec SqueezeNet V1.1 sur une image prise d'internet [1].

Les méthodes présentées précédemment permettent de générer des cartes de saillance mettant en évidence les régions les plus intéressantes dans l'image par le CNN. Ces méthodes ont été combinées avec les CNN en proposant une couche d'attention qui est entraînable. Une telle couche permet de capturer les régions spatiales des caractéristiques qui sont plus intéressantes et les prioriser au reste des régions [94].

Dans le cadre de nos travaux, nous proposons un mécanisme d'attention basé sur les cartes de saillance pour avoir une compréhension visuelle de la décision prise. Grâce à ce mécanisme d'attention, une analyse plus profonde des caractéristiques spatio-temporelles apprises est faite afin de justifier la prise de décision.

1. source de l'image du chien avec le chat : https://urlz.fr/g1dR

5.5.1 Mécanismes d'attention

Dans cette section, nous présentons deux mécanismes d'attention dans lesquels nous utilisons les méthodes des cartes de saillance sur nos STR afin de visualiser d'un coté la partie spatiale de la courbe Γ utilisée pour créer la STR la plus discriminante et d'un autre coté la plage temporelle la plus significative pour la classification.

Dans ce contexte, nous proposons un mécanisme d'attention basé sur ces cartes de saillances par lequel nous étudions les deux domaines séparément. Une attention temporelle est proposée quand seulement le domaine temporel est étudié et une attention spatiale est proposée permettant la génération d'une carte de segmentation sémantique en étudiant seulement l'information spatiale portée par les STR. Dans cette étude, nous utilisons les cartes de saillances S^c générées avec la méthode $GradCAM++$ [21]. La méthode choisie est moins coûteuse que $ScoreCAM$ [137] car ce dernier doit faire passer l'image plusieurs fois au CNN pour avoir le vecteur des scores lui permettant de raffiner la carte S^c.

5.5.1.1 Attention temporelle

L'attention consiste ici à analyser le domaine temporel le plus discriminant pour la décision. Par le biais de cette étude, nous pouvons localiser la plage temporelle la plus discriminante en fonction des classes étudiées puis vérifier si le modèle peut fournir de meilleurs scores sur cette nouvelle plage temporelle.

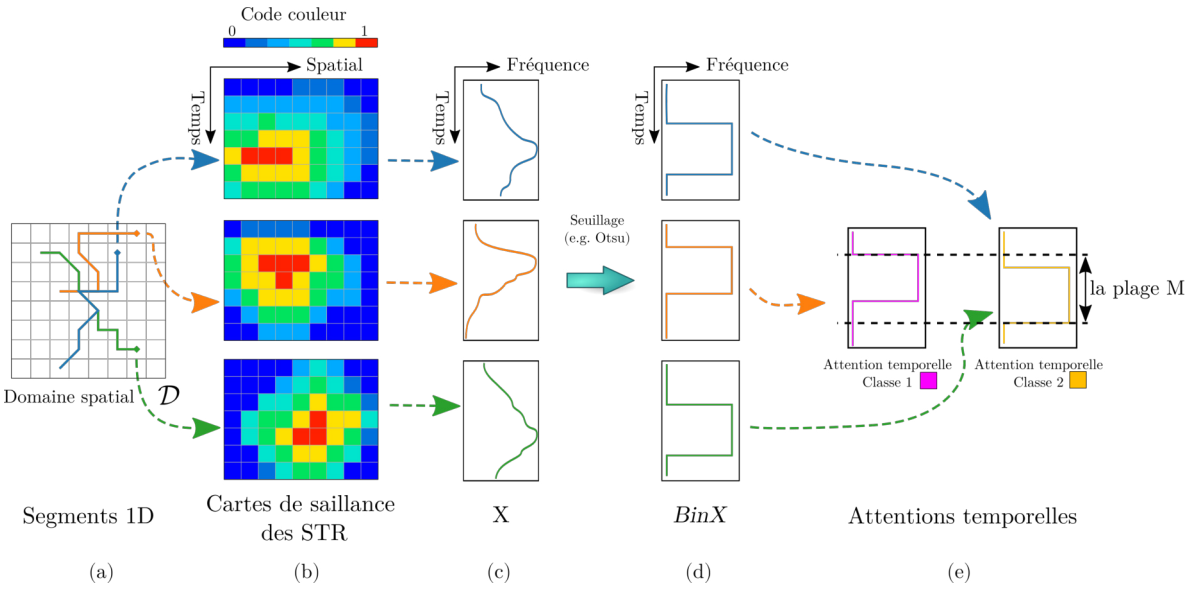


Figure 5.12 – Attention temporelle dans l'approche $MS-STR$: (a) Les segments RW (Les points dans les courbes sont leurs débuts) ; (b) Les cartes de saillance des N_{seg} STR ; (c) Les profils d'attention temporelle ; (d) Les profils binarisés ; (e) Le masque M pour l'attention temporelle globale.

Dans un premier temps, nous rappelons qu'une STR contient l'information temporelle en vertical et l'information spatiale en horizontal. La carte de saillance S^c associée à la STR contient donc également l'information temporelle en vertical et l'information spatiale en horizontal. Pour ne considérer que l'attention temporelle, nous nous focalisons donc sur l'axe vertical.

Soit une STI donnée, N_{seg} STR lui sont associées. Nous calculons alors une carte de saillance S^c pour chaque STR. Notons que les cartes de saillance S^c sont considérées comme des images d'attention de taille $T \times L$. La figure 5.12.b illustre trois cartes S^c associées aux segments présentés dans la figure 5.12.a.

Nous présentons maintenant le processus de l'attention temporelle pour une STR. Soit une STR, C cartes de saillance (S) sont obtenues, une pour chaque classe. Dans notre cas, nous limitons le choix de la carte S à la classe $c(STR)$ prédite par le CNN. La carte S^c obtenue est interprétée de la manière suivante, chaque valeur $S_{i,j}^{c(STR)}$ désigne l'attention à la date $i \in [1, T]$ (ligne) et au pixel $j \in [1, L]$ (colonne). Dans cette partie, le but est d'analyser l'attention temporelle seulement. Pour ce faire, nous commençons par définir les profils d'attention temporelle, notés X, pour chaque STR. La taille du profil X est T. Le profil est obtenu en cumulant les valeurs sur les lignes de S^c. L'équation 5.7 présente la formule de l'accumulateur :

$$X_i(STR) = \sum_{j=1}^{L} S_{i,j}^{c(STR)}, \quad \forall i \in [1, T] \tag{5.7}$$

où X_i est la i^{ime} coordonnée du vecteur X et $X(STR)$ est le vecteur d'attention temporelle associé à une STR. La figure 5.12.c illustre les profils d'attention temporelle obtenus pour chaque STR. À partir de l'analyse de tous les profils temporels associés aux éléments d'apprentissage, nous pouvons sélectionner la plage temporelle la plus discriminante de chaque classe que le CNN a trouvé. Pour ce faire, nous commençons par capter l'attention temporelle la plus forte pour chaque STR. Cela est fait en appliquant un seuillage Bin pour que les profils X soient définis dans $\{0, 1\}$, les profils binarisés obtenus sont notés $BinX$. Les valeurs élevées dans X sont celles égales à 1 dans $BinX$ et inversement. Toutefois, le seuillage Bin nécessite un seuil qui peut être fourni par un expert ou peut être trouvé de façon automatique. Dans notre cas, nous utilisons la méthode d'Otsu pour définir le seuil de la fonction Bin de manière globale. La figure 5.12.d illustre les résultats obtenus.

Les profils binarisés ont permis de capter l'attention temporelle la plus forte pour chaque STR. Il est possible de généraliser cette attention en construisant un masque M qui représente la plage temporelle globale que le modèle a trouvé pour séparer les classes. Le calcul du masque M est basé sur les vecteurs $BinX$. Pour cela, nous calculons le produit scalaire des $BinX^c$. Comme chaque STR est labellisée parmi une des C classes et que le modèle est entraîné à prédire le même label, $BinX$ met en évidence la plage temporelle la plus discriminante pour cette STR. La figure 5.12.e illustre les profils temporels obtenus pour chaque classe. Maintenant, au lieu d'utiliser tout le domaine temporel, nous pouvons le réduire seulement aux dates auxquelles la valeur $(BinX^c)_i$ est égale à 1. L'équation 5.8 permet la

création du masque M :

$$M = (M_i)_i = \begin{cases} 1 & \text{si} \quad \exists k \in [1, C] \quad \prod_{STR/c(STR)=k} BinX_i^{c(STR)} = 1 \\ 0 & \text{sinon} \end{cases} \tag{5.8}$$

5.5.1.2 Attention spatiale

L'attention spatiale a pour but d'analyser l'information spatiale qui a été utile pour le CNN.

Nous faisons l'hypothèse que S^c met en évidence les régions spatiales de l'image qui contribuent davantage à la prise de décision. En appliquant S^c sur les STR, nous pouvons avoir une information visuelle sur l'information spatiale portée par la courbe Γ la plus discriminante pour le CNN. En reprojetant les valeurs d'attention de S^c sur le domaine original $\mathcal{D}$ du support $2D$ de la STI, nous pouvons interpréter visuellement la décision prise par le CNN conduisant à produire une segmentation sémantique à partir des cartes de saillance S^c.

Dans une STR, l'information spatiale est localisée sur l'axe horizontal. Suivant la même stratégie que pour l'attention temporelle, pour chaque pixel nous considérons sa colonne dans la carte de saillance S^c. En revanche, au lieu de faire la somme des valeurs d'attention,

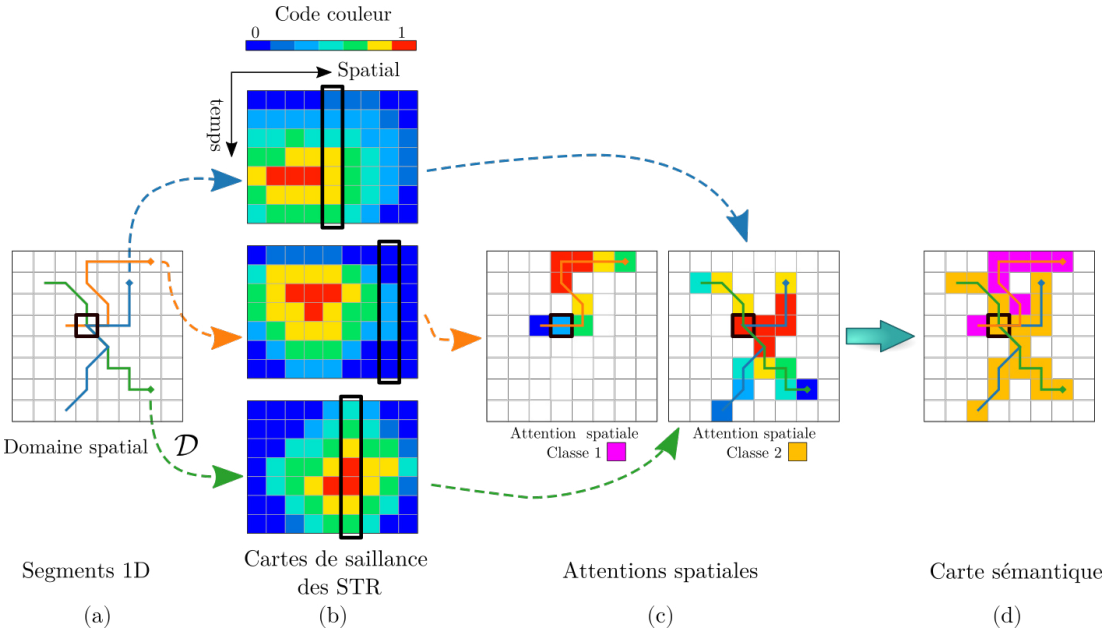


Figure 5.13 – Attention spatiale dans l'approche $MS-STR$: (a) Les segments RW (Les points dans les courbes sont leurs débuts) ; (b) Les cartes de saillance des N_{seg} STR ; (c) La rétro-projection des valeurs d'attention dans le domaine spatial $\mathcal{D}$ de l'image ; (d) Le résultat de la carte sémantique.

nous considérons le moment où le pixel est le plus attractif pour la classe c. Un pixel peut avoir une attention à un instant donné ou sur une plage temporelle. En faisant la moyenne des valeurs d'attention, certains pixels perdront de leur valeur car ils étaient à leur attention maximale à des courts moments par rapport à d'autres. Par contre en prenant le maximum, nous conservons la valeur de l'attention. La figure 5.13.b présente des cartes de saillance pour différentes STR. Les deux cartes S^c du bas montrent un exemple où un pixel peut avoir une ou plusieurs valeurs d'attention maximales. Pour cela, capter uniquement l'attention la plus forte empêchera de perdre les valeurs d'attention maximales pour tous les pixels. Pour ce faire, nous définissons le vecteur Y^c comme :

$$Y^c_j = Y^c_{(x,y)} = \max_{i=1}^{N} S^c_{i,j} \tag{5.9}$$

où Y^c_j est l'attention maximale associée au pixel d'indice j dans la STR et de coordonnées (x, y) dans $\mathcal{D}$, le domaine de l'image originale.

Suivant les approches globales $G - STR$ et locales $MS - STR$, le processus de reprojection ne peut être le même. Dans l'approche globale $G - STR$, une seule image peut être construite car nous trouvons une seule valeur $Y^c_{(x,y)}$ pour chaque point de coordonnées (x, y). Dans l'approche locale $MS - STR$, le processus de reprojection devient plus complexe. La STI est représentée par un ensemble de N_{seg} STR qui ont la même longueur L, notés STR^k où k est inclus dans l'intervalle $[1, N_{seg}]$. Un pixel de coordonnées (x, y) défini dans $\mathcal{D}$ peut figurer dans plusieurs STR. Ce dernier a donc différentes valeurs d'attention $Y^{c(STR^k)}_{(x,y)}$ en fonction des S^c associées à chaque STR^k et sa classe $c(STR^k)$ prédite par le CNN. Mais, nous ne gardons ici que l'attention maximale pour ce pixel parmi toutes les valeurs obtenues. Pour ce faire, l'attention spatiale d'un pixel est calculée avec la formule 5.10 définie ci-dessous :

$$Y^c_{(x,y)} = \max\left(Y^{c(STR^k)}_{(x,y)} \; k \in [1, N_{seg}] \text{ et } c(STR^k) = c, \;\; 0\right) \tag{5.10}$$

La figure 5.13.a illustre un pixel appartenant à trois courbes, deux étiquetées c_1 et une étiquetée c_2. Leurs trois cartes de saillance S^c sont présentées dans la figure 5.13.b. Dans chaque cas, le pixel appartenant aux trois courbes est encadré avec un rectangle dans les cartes S^c. Nous ne considérons que la valeur la plus élevée de l'attention spatiale de ce pixel pour chaque STR (équation 5.9). Enfin pour chaque classe c, nous gardons les valeurs d'attention les plus hautes parmi celles de toutes les STR. La figure 5.13.c illustre l'attention spatiale concernée pour les deux classes. La classe c_2 montre le cas où la valeur d'attention maximale est conservée entre deux STR.

Pour obtenir une carte sémantique, nous construisons une carte couvrant le domaine spatial $\mathcal{D}$ de la STI. Pour ce faire, l'étiquette du pixel est celle où la valeur d'attention est la plus grande parmi les classes possibles :

$$V_{(x,y)} = \arg\max_{c \in [1,C]} \max_{k \in [1,N_{seg}]/c=c(STR^k)} Y^{c(STR^k)}_{(x,y)} \tag{5.11}$$

La figure 5.13.d présente le résultat de la carte de segmentation sémantique pour les deux classes.

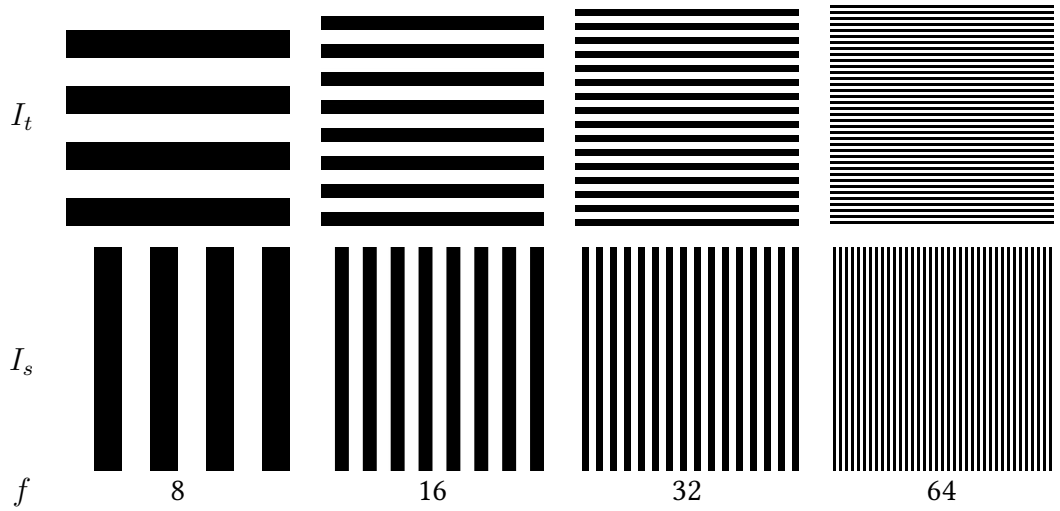


Figure 5.14 – Images synthétiques obtenues avec les différentes valeurs de f.

5.5.2 Nature des informations impliquées dans la décision

Grâce aux informations spatiale et temporelle portées par les STR, les CNN extraient des caractéristiques spatio-temporelles avec des convolutions $2D$. Dans la suite, nous analysons les résultats de classification obtenus avec les STR afin d'évaluer le type d'information extrait avec le CNN.

Pour ce faire, nous proposons d'analyser les filtres de convolution qui ont été appris lors de l'entraînement du CNN. Les premières couches du CNN contiennent des couches convolutives. En fonction de la nature des données, les STR, de telles couches traitent en vertical l'aspect temporel et en horizontal l'aspect spatial. Par contre, il est difficile d'expliquer quelle est l'information (temporelle, spatiale ou spatio-temporelle) qui a été la plus utilisée. Dans ce contexte, nous proposons une stratégie d'analyse des filtres qui permet d'identifier le domaine traité par chacun des filtres. Pour cela, nous construisons d'abord des images synthétiques contenant seulement une information temporelle I_t ou seulement une information spatiale I_s. Les images synthétiques utilisées I_s et I_t sont générées en alternant deux couleurs (noir et blanc) f fois où f est le nombre de bandes. Dans notre cas, nous avons fait varier f afin d'étudier l'adaptation des filtres aux changements. Les valeurs de f sont 8, 16, 32 et 64. La figure 5.14 illustre les images I_s et I_t synthétiques obtenues avec les différentes valeurs de f. Ces fausses images sont fournies comme entrée au CNN, après l'entraînement de celui-ci sur un problème donné. La prochaine étape consiste à calculer, pour une couche donnée, l'énergie des caractéristiques F, obtenues par les convolutions, de chaque filtre (indice k), notée $E_k(F_t)$ et $E_k(F_s)$. Un ratio est ensuite calculé qui est un rapport spatio-temporel $R_{st}(k)$ entre les deux énergies. Ce ratio indique quel aspect (spatial ou temporel) est plus ou moins associé au filtre k :

$$R_{st}(k) = \frac{E_k(F_s)}{E_k(F_t)} \tag{5.12}$$

Enfin, la nature des filtres (spatiale, temporelle ou spatio-temporelle) peut être définie en définissant deux seuils μ et ν. Ces derniers sont des paramètres de la méthode. En fonction de μ et ν, nous pouvons considérer trois types de filtres :

- **Les filtres spatiaux** sont ceux qui sont plus liés aux variations spatiales : le rapport $R_{st}(k)$ est supérieur à $1 + \mu$;
- **Les filtres temporels** sont ceux qui sont plus liés aux variations temporelles : le rapport $R_{st}(k)$ est inférieur à $1 - \nu$;
- **Les filtres spatio-temporels** sont ceux dont le rapport $R_{st}(k)$ est compris entre $1-\nu$ et $1 + \mu$; ils sont liés à la fois aux variations temporelles et spatiales.

5.6 Étude expérimentale

Nous présentons maintenant l'étude expérimentale que nous avons menée afin d'évaluer la méthode $Deep - STaR$ dans nos deux cadres applicatifs. La première application concerne le domaine de la télédétection. Elle consiste à analyser la couverture des sols afin d'aider les décideurs politiques en matière d'agriculture et d'environnement. La deuxième application se focalise sur le problème de détection de violence dans les vidéos.

Les données de vérité terrain utilisées sont celles présentées dans le chapitre 3. Dans la suite, nous expliquons les détails des préparations des STR dans la section 5.6.1. La section 5.6.2 présente les protocoles de validation des expérimentations. Enfin, les sections 5.6.3 et 5.7 exposent respectivement les résultats obtenus et le bilan scientifique.

5.6.1 Préparation des STR

Tout d'abord, nous commençons par préparer les STR associées à chacune des méthodes $G - STR$ et $MS - STR$. Les CNN classiques $2D$ dédiés à la classification traitent des images de taille 224×224. Nous devons donc adapter les images des STR générées à cette taille dans les deux dimensions, temporelle et spatiale. Nous expliquons par la suite comment nous réalisons cette tâche.

5.6.1.1 Dimension temporelle (axe vertical)

Afin d'adapter l'axe temporel des STI à la hauteur de la STR, deux stratégies sont possibles. La première stratégie consiste à considérer directement la STR si $T = 224$. Dans le cas où $T \leq 224$, nous remplissons avec des zéros jusqu'à atteindre les 224 valeurs tout en centrant les T valeurs disponibles. Par contre, dans la deuxième stratégie, nous appliquons une interpolation linéaire sur la dimension temporelle pour générer 224 dates à partir des T valeurs initiales.

Dans l'application de télédétection, nous utilisons les deux stratégies dans l'analyse des parcelles agricoles car nous supposons que l'évolution de la végétation est monotone et linéaire entre deux dates consécutives. Par contre pour la reconnaissance de la violence à partir de vidéos, nous n'utilisons que la première stratégie car la deuxième influe sur la vitesse de déplacement des objets dans la scène ce qui change la réalité du mouvement observé.

5.6.1.2 Dimension spatiale (axe horizontal)

Concernant la dimension spatiale, l'adaptation peut être faite de différentes manières en accord avec les deux stratégies, locale $MS-STR$ et globale $G-STR$.

- **Stratégie locale** $MS-STR$ **:** comme expliqué dans la section 5.2.2.2, la stratégie locale permet de générer un nombre élevé N_{seg} de STR et ce pour chaque STI. Grâce au processus RW, les segments peuvent capturer différentes configurations spatiales par STI c'est-à-dire en passant par un même pixel dans différentes directions dans les représentations $2D$. Afin d'analyser la qualité de l'information spatiale enrichissant les séries temporelles de pixels, nous faisons varier la longueur L des segments. Les longueurs étudiées sont 10, 50 et 100. Enfin, nous centrons les représentations planaires obtenues sur l'axe horizontal de l'image d'entrée du CNN et nous fixons des valeurs nulles pour les autres pixels de l'image STR d'entrée ;
- **Stratégie globale** $G-STR$ **:** cette stratégie permet de créer une seule STR par STI. Toutefois, le nombre de pixels dans le domaine spatial $\mathcal{D}$ d'une STI n'est pas toujours égal à 224. En fonction du nombre de pixels d'une STI, deux stratégies permettent d'avoir 224 colonnes : (1) si il y a moins de 224 pixels, nous répétons la séquence jusqu'à ce que les 224 valeurs soient remplies ; (2) sinon, si le nombre de pixels est supérieur à 224, nous divisons la séquence en différentes images de 224 valeurs. Pour la deuxième stratégie, nous utilisons la même méthode de prise de décision que quand une STI est représentée par plusieurs STR comme expliqué dans la section 5.4.

Les CNN ont besoin d'un ensemble d'apprentissage. Ce dernier permet d'optimiser les paramètres constituant un modèle. Par contre, les CNN sont toujours confrontés à des problèmes de généralisation. Les deux problèmes principaux sont le sous-apprentissage et le sur-apprentissage. Le premier problème concerne l'incapacité du modèle à s'optimiser. Quant au sur-apprentissage, le modèle apprend par cœur les données d'entraînement et il sera dans l'incapacité de faire des prédictions correctes sur de nouvelles données. Afin d'éviter tous ces problèmes, il est important que l'ensemble d'entraînement contienne un grand nombre d'images. Pour cela, nous expliquons dans la suite la quantité des STR que nous allons générer pour chacune des applications.

Nombre de STR pour l'analyse des $STIS$

Pour l'application de télédétection, les aires des parcelles agricoles à classifier sont très limitées. Pour la stratégie locale $MS-STR$, nous avons fait varier le nombre N_{seg} de

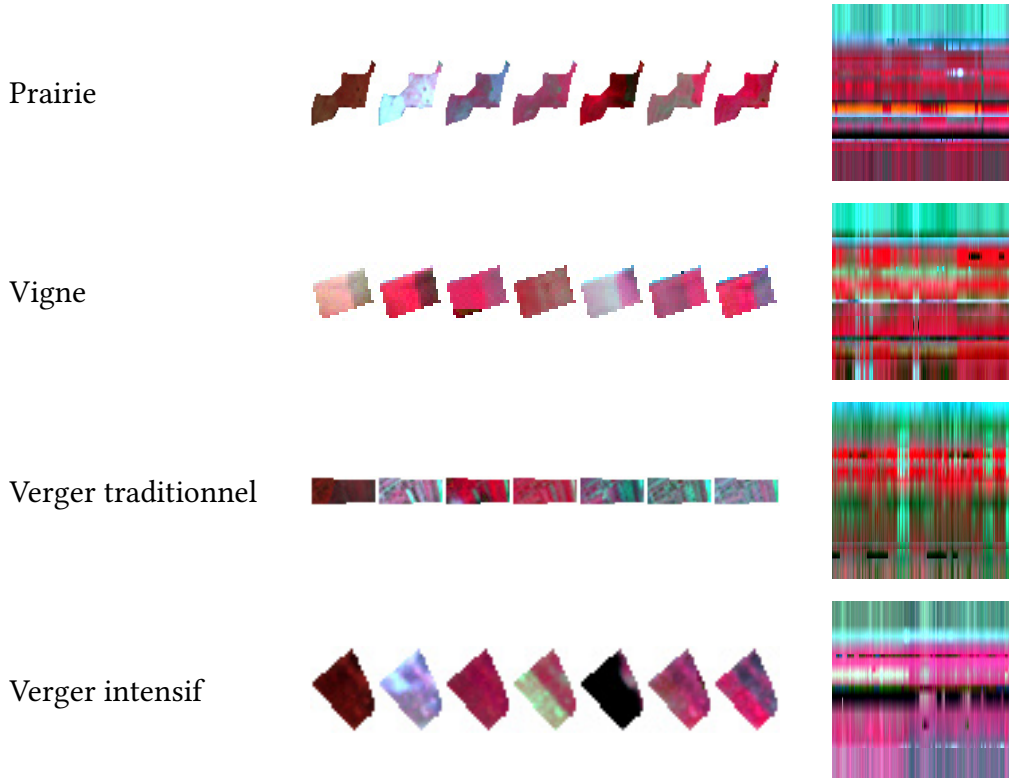


Figure 5.15 – Évolution temporelle de quatre parcelles agricoles des différentes classes thématiques étudiées. (À gauche) la représentation $2D+t$ des données $STIS$; (À droite) leurs STR crées avec la méthode $MS-STR$. La longueur du segment est 224.

Tableau 5.1 – Nombre de STR générées pour les deux approches $MS-STR$ et $G-STR$ pour l'application de télédétection. Les pourcentages sélectionnés sont les mêmes pour l'apprentissage et le test.

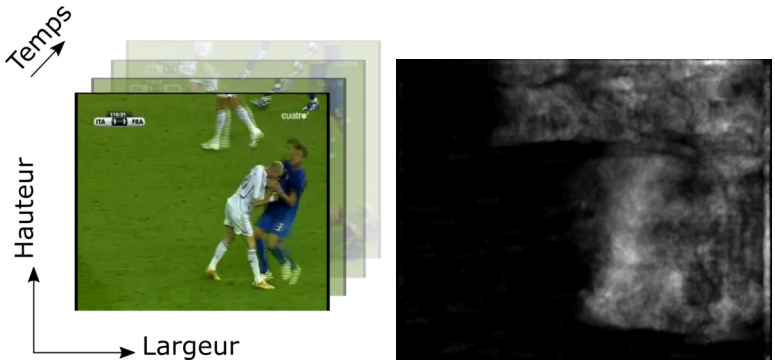
Classes	**# MS-STR**				**# G-STR**
	$RW_{10\%}$	$RW_{20\%}$	$RW_{50\%}$	$RW_{70\%}$	$\Re_*$
Prairies	26 110	51 688	128 424	179 914	1 757
Vignes	3 060	5 821	14 137	19 853	577
Vergers traditionnels	2 146	4 222	10 474	14 672	189
Vergers intensifs	2 564	5 027	12 414	17 408	226
Total	33 880	66 758	165 449	231 847	2 749

segments en fonction de la surface de chaque parcelle, c'est-à-dire le nombre de pixels la composant. Le N_{seg} utilisé pour constituer l'ensemble d'apprentissage est le même que celui du test. Nous utilisons expérimentalement 10%, 20%, 50% et 70% de la taille de chaque parcelle. Par contre pour la stratégie globale $G-STR$, seules quelques parcelles vont avoir

plus qu'une STR. Le tableau 5.1 résume le nombre de STR générées pour les deux stratégies. Nous remarquons l'existence d'une différence élevée dans le nombre de représentations générées entre les deux stratégies. Nous proposons alors d'utiliser une technique d'augmentation de données pour la stratégie globale $G - STR$ qui sera détaillée dans la section 5.6.1.3. La figure 5.15 illustre quelques exemples de parcelles agricoles.

Nombre de STR pour l'analyse des vidéos

Concernant l'analyse des vidéos, globalement chaque vidéo a sa propre étiquette. Mais au niveau spatial, la violence est localisée dans une certaine région qui peut évoluer au cours du temps. Pour prendre en compte ce phénomène de manière à construire des STR correctement étiquetées pour l'apprentissage, nous commençons par définir une région globale d'intérêt représentant la violence en localisant les zones les plus mouvementées. La méthode utilisée pour réaliser cette tâche est inspirée de celle utilisée dans [22]. Elle est basée sur le flux optique dense. Nous avons sélectionné la méthode de Gunnar Farneback [36] qui est une extension de celle de Lucas-Kanade. Tout d'abord, nous calculons le flux optique de toute la vidéo. Puis, nous calculons la moyenne du flux obtenu sur l'axe temporel.



(a) Vidéo originale (b) Flux moyen

(c) Binarisation du flux moyen (d) ROI obtenue après post-traitement

Figure 5.16 – Illustration des différentes étapes de l'extraction de la région violente dans une vidéo.

Enfin, la région d'intérêt, notée ROI comme *Region Of Interest*, est obtenue en appliquant une binarisation sur le flux moyen pour obtenir la région spatiale à fort mouvement que nous supposons représenter une zone violente. Dans notre cas, nous utilisons la méthode Otsu pour choisir automatiquement un seuil optimisé dans le flux moyen. Un post-traitement est appliqué afin de ne conserver qu'une région globale (une seule composante connexe). La figure 5.16 illustre cette étape.

Concernant la création des STR, nous nous limitons à la stratégie locale $MS-STR$ car la violence est localisée spatialement. En effet, si nous appliquons la stratégie globale, nous serions dans l'obligation de diviser chaque STR en plusieurs STR qui n'auront pas la même étiquette. Cette stratégie revient à une représentation multi-segments. Le nombre N_{seg} de segments ne peut pas être fixé de la même façon qu'avec les $STIS$ car les tailles des vidéos sont largement supérieures à celles des parcelles. Pour cela nous utilisons un N_{seg} fixe pour la phase d'apprentissage et nous le faisons varier lors de la prise de décision.

Pour le nombre de STR, nous avons choisi empiriquement N_{seg} égal à 30 par vidéo. Les vidéos violentes contiennent à la fois la violence qui est subie dans les ROI et de la non violence. Cependant, la non violence peut être présente partout dans les vidéos non violentes. Pour cela, nous générons 30 STR violentes dans les ROI des vidéos violentes. Par contre pour les STR non violentes, nous avons choisi d'en générer 28 dans les vidéos non violentes et 2 hors ROI dans les vidéos violentes. Une telle stratégie permet d'avoir un nombre équilibré d'instances N_{seg} pour chaque classe. Le tableau 5.2 résume le nombre de STR générées pour les vidéos des différentes bases. Pour l'ensemble de test, N_{seg}^{test} STR sont générées sans utiliser la ROI. Pour prendre une décision, nous avons pris différentes valeurs pour N_{seg}^{test} : 10, 30, 50 ou 100.

Tableau 5.2 – Nombre de STR générées pour les ensembles d'entraînement des différentes bases de vidéos.

Méthode	***RWF2000***	***Movies fights***	***Hockey fights***	***Crowd Violence***
Vidéo violente	30 000	3 000	15 000	3840
Vidéo non violente	30 000	3 000	15 000	3840
Total	60 000	6 000	30 000	7680

5.6.1.3 Augmentation des données pour l'approche globale

Dans le tableau 5.1, nous remarquons que les ensembles de données pour l'approche globale $G-STR$ sont relativement plus petits par rapport à ceux de l'approche locale $MS-STR$. Cela est dû aux multiples représentations STR par STI basées sur les RW. Puisque nous utilisons une méthode d'apprentissage profond pour apprendre les caractéristiques spatio-temporelles et classifier les STR, l'approche globale $G-STR$ est désavantagée car l'entraînement d'un modèle sur peu de données a tendance à sur-apprendre en raison de la

quantité limitée de données. Pour faire face à cette limite, nous avons étudié des techniques utilisées pour éviter ce phénomène de sur-apprentissage.

Les premières techniques proposées dans la littérature sont basées sur des fonctions de généralisation comme le *dropout* [125] ou la normalisation par lot [61]. D'autres techniques transforment directement les données d'entrée afin d'en avoir d'autres qui sont générées de façon synthétique. Ces techniques sont dites techniques d'augmentation des données (AD). Ces dernières se divisent en deux groupes. Le premier groupe représente les méthodes traditionnelles basées sur des transformations d'images. Les plus courantes sont les transformations affines telles que le retournement, la rotation et la translation. Il existe également des transformations non affines telles que le redimensionnement, le recadrage, le rognage ou l'ajout de bruit. Le deuxième groupe concerne les méthodes les plus récentes qui sont basées sur des réseaux antagonistes génératifs, en anglais *Generative Adversarial Network* (GAN), pour produire des données synthétiques supplémentaires [121]. Dans notre cas, de telles méthodes ne peuvent pas être considérées pour générer plus de données puisque nous avons besoin d'une énorme base de données annotées qui n'est pas disponible dans le contexte de ces études thématiques. Dans ces dernières, nous utilisons uniquement des transformations affines classiques appliquées sur le domaine spatial initial $\mathcal{D}$ afin de conserver la résolution spatiale réelle des données. Nous avons choisi d'appliquer des rotations avec les angles suivants : $45°$, $90°$, $135°$ et $180°$.

5.6.1.4 Normalisation et standardisation des données

La normalisation ou la standardisation des données est une étape importante dans toute application de classification. Cette transformation a pour but la mise en échelle des données afin qu'elles soient comparables entre elles. Un tel traitement des données joue un rôle crucial dans les tâches de reconnaissance des formes. Pour ce faire, deux types de transformations sont possibles. La première est dite normalisation et la deuxième standardisation.

La normalisation est une méthode traditionnelle qui a pour but de ramener le domaine des données au domaine $[0, 1]$. Différentes méthodes de normalisation existent. La plus courante est basée sur les valeurs maximale et minimale d'un ensemble X de données. La formule 5.13 permet le calcul de la normalisation :

$$x' = \frac{x - min(X)}{max(X) - min(X)}, \quad \forall x \in X \tag{5.13}$$

La standardisation est davantage utilisée dans les problèmes de classification d'images. Elle est aussi connue sous le nom de Z-normalisation. Elle consiste à soustraire de chaque donnée la moyenne de l'ensemble X des données, puis à la diviser par l'écart-type de X. Une telle transformation conduit à ce que les valeurs des données aient une moyenne nulle et une variance égale à 1. La formule 5.14 permet le calcul de la standardisation d'un échantillon de l'ensemble X :

$$x' = \frac{x - \mu(X)}{\sigma(X)}, \quad \forall x \in X \tag{5.14}$$

Dans notre cas, nous utilisons la normalisation pour analyser les STR des $STIS$ pour ne pas changer l'évolution temporelle des données. Cependant, nous avons limité les valeurs minimum et maximum respectivement à 2% et 98% des percentiles [98] afin d'éliminer les intensités extrêmes qui sont dues à des problèmes naturels, comme les nuages. Par contre nous utilisons la standardisation pour analyser les STR de vidéos.

5.6.2 Protocole de validation

Nous avons utilisé un protocole de validation croisée dans les deux expérimentations qui sont la classification de parcelles agricoles et la classification de vidéos. Tout d'abord, nous divisons les données en trois ensembles associés à l'entraînement, à la validation et au test. Ils ont respectivement les proportions suivantes : 60%, 20% et 20% du nombre de STI total. Nous indiquons que les ensembles sont stratifiés c'est-à-dire qu'ils respectent les proportions des classes. Le modèle est entraîné cinq fois et nous donnons le taux de reconnaissance moyen et l'écart-type. Par contre, dans le cas où nous traitons la base de vidéo *RWF2000*, nous extrayons seulement un ensemble de validation de celui d'entraînement et nous testons directement sur l'ensemble de test proposé par les créateurs de la base [22] de manière à rendre possible une comparaison des résultats.

Concernant l'entraînement, nous avons utilisé comme fonction de perte la *cross entropy*. Pour l'application de classification des parcelles, nous pondérons la fonction de perte afin de traiter le problème des classes déséquilibrées en évitant le sur-apprentissage vers les classes majoritaires. Les poids de pondération de la fonction de perte sont basés sur la fréquence d'apparition de chaque classe dans la base globale.

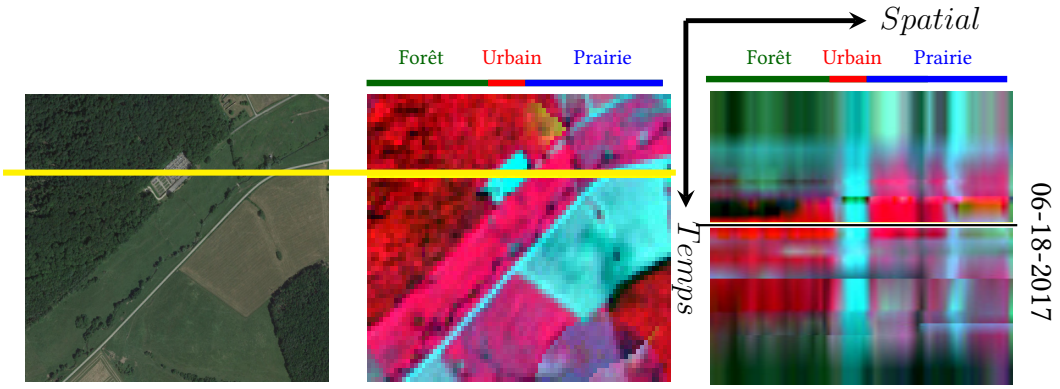
L'optimiseur utilisé est *Adam* avec un taux d'apprentissage de 10^{-6} pour la classification des parcelles et 10^{-5} pour la classification des vidéos. Nous gardons le reste des paramètres par défaut ($\beta_1 = 0.9, \beta_2 = 0.999$ et $\epsilon = 10^{-8}$). L'apprentissage est arrêté avec la technique d'arrêt précoce (*early stopping*) avec un nombre de patience de 10. Les expériences sont réalisées sur un serveur équipé d'un GPU Nvidia Tesla T4.

Dans cette étude, l'entraînement du modèle est réalisé avec deux stratégies, soit à partir d'une initialisation aléatoire des poids, soit initialisé avec les poids obtenus quand le modèle est entraîné sur la base de ImageNet dans une tâche de classification (*i.e.*, le challenge de ILSVRC 2012 [2]) et ensuite affiné avec nos données.

5.6.3 Résultats et discussions

Suite à la présentation du protocole, nous allons maintenant présenter puis discuter des résultats de la classification sur l'application de télédétection, suivi des résultats de la classification des vidéos de violence. En télédétection, nous comparons les approches locale

2. ImageNet Large Scale Visual Recognition Challenge 2012 : https://www.image-net.org/challenges/LSVRC/2012/index.php



(a) Image de *Google Earth* (1665 × 2056 pixels)

(b) Image Sentinel-2 prise le 06-18-2017 (62 × 78 pixels)

(c) STR associée au segment jaune dans (b)

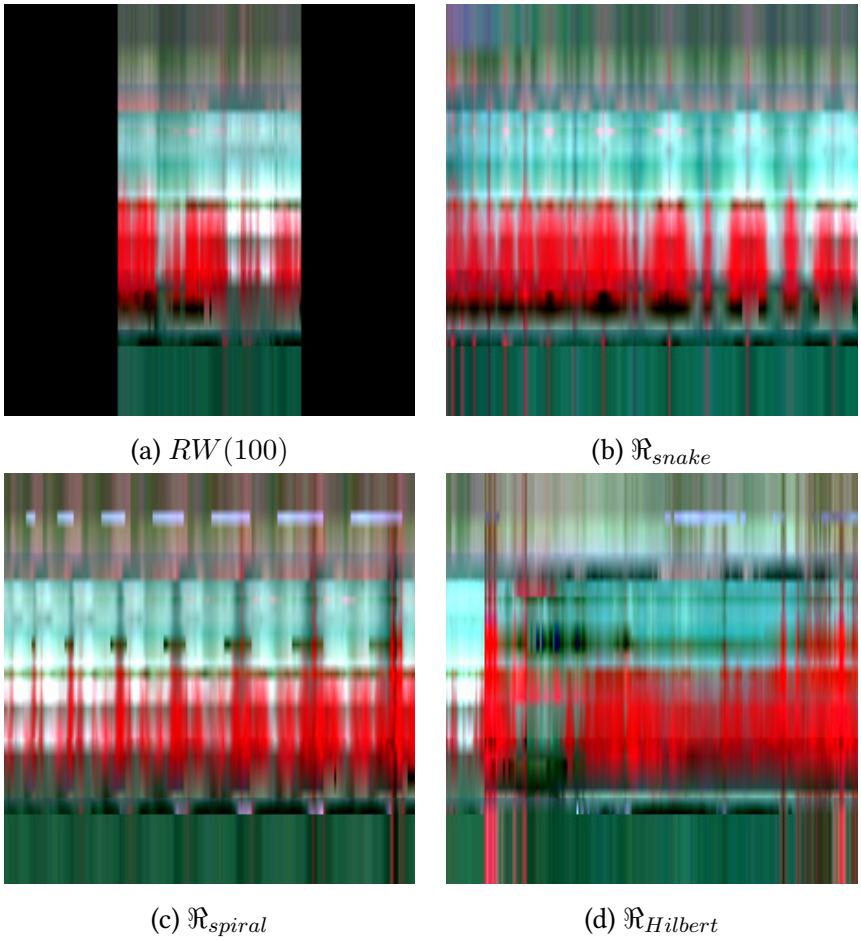
Figure 5.17 – Un exemple d'une STR sur une $STIS$ de Sentinel-2 : (a) Une image à très haute résolution spatiale prise de *Google Earth* sur une zone agricole particulière ; (b) Une image Sentinel-2 prise le 18 juin 2017, cette image appartient à une $STIS$ prise au cours de l'année 2017 sur la même région que l'image *Google Earth* ; (c) La représentation spatio-temporelle créée à partir du segment jaune dans l'image Sentinel-2.

$MS-STR$ et globale $G-STR$ mais nous nous limitons à l'approche locale $MS-STR$ pour la détection de violence. Nous présentons également quelques comparaisons avec des méthodes de l'état de l'art concurrentes dans les deux applications.

5.6.3.1 Application en télédétection

De manière à mieux comprendre la nature des STR et à interpréter le contenu des images associées, nous présentons sur la figure 5.17 présente (a) une image haute résolution de Google Earth d'une zone agricole spécifique, (b) une image Sentinel-2 de la même région acquise le 18 juin 2017 ainsi (c) qu'une STR correspondant à un chemin (segment jaune) simple traversant la zone. Le chemin passe par différentes zones, sur la partie gauche se trouve une forêt, puis au milieu un bâtiment et enfin à droite deux prairies qui ne sont pas du même type. La STR créée à partir de ce chemin est présentée dans l'image (c). Dans cette représentation, le segment noir montre la date à laquelle l'image Sentinel-2 en (b) a été acquise. Les discontinuités dans le sens vertical peuvent correspondre à la présence des nuages non détectés.

Dans la partie gauche de la représentation, on peut observer qu'une zone assez homogène sur la direction horizontale est associée à la zone forestière. Verticalement, la zone rouge indique la présence de chlorophylle active de juin à octobre. Ensuite, au milieu, un rectangle bleuté stable tout le long de l'année et étroit est associé à la maison avec une faible valeur dans le proche-infrarouge (NIR) qui est plus généralement lié à la non présence de la végétation. Dans la partie droite, nous distinguons deux comportements, une première prairie avec un intervalle temporel rouge indiquant que l'herbe est en train de pousser et la



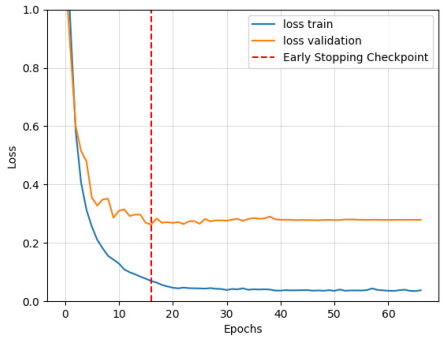
(a) $RW(100)$ (b) $\Re_{snake}$

(c) $\Re_{spiral}$ (d) $\Re_{Hilbert}$

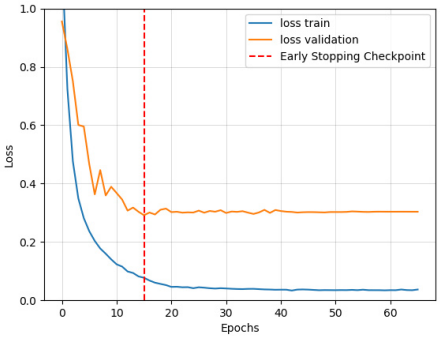
Figure 5.18 – Illustration des STR obtenues avec les deux approches $MS-STR$ et $G-STR$; (a) STR de $MS-STR$ avec un Random Walk de $L=100$ ($RW(100)$). 62 colonnes noires sont rajoutées des deux cotés afin d'obtenir une image de 224×224; (b, c, d) STR de $G-STR$ générées avec les différentes courbes remplissant l'espace.

seconde prairie où la croissance de la végétation se produit pendant une période temporelle plus courte. Cela peut indiquer que la prairie a été cultivée à la fin du printemps et fauchée, la végétation est alors arrêtée et une zone bleue est présente. Nous pouvons également remarquer que l'herbe pousse avant les feuilles des arbres. La route entre des deux prairies est associée à la zone bleue stable tout au long de l'année.

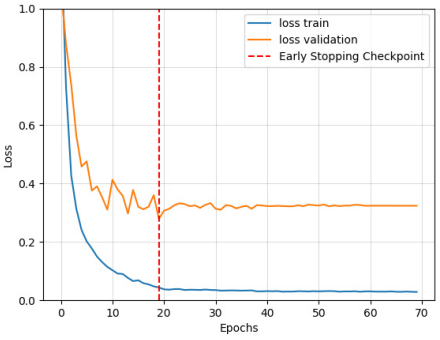
La figure 5.18 affiche quelques STR d'un verger traditionnel. Ces dernières sont générées avec les deux approches, locale $MS-STR$ et globale $G-STR$. Du coté de l'approche globale $G-STR$, nous présentons une STR pour chaque courbe remplissant l'espace. Avec $\Re_{snake}$, des motifs réguliers apparaissent horizontalement tandis qu'avec $\Re_{spiral}$, les motifs deviennent plus grands de gauche à droite car la courbe commence au centre de l'image et va plus loin avec une plus grande amplitude. Comme prévu, $\Re_{Hilbert}$ conduit à un résultat plus lisse que les autres car cette courbe préserve mieux la localité des pixels. Pour l'approche locale, RW fournit un résultat lisse comme le $\Re_{Hilbert}$.



(a) $\Re_{snake}$



(b) $\Re_{spiral}$



(c) $\Re_{Hilbert}$

Figure 5.19 – Courbes de pertes obtenues quand le modèle est entraîné sur les STR de $\Re_{snake}$, $\Re_{spiral}$ et $\Re_{Hilbert}$ pour l'application de télédétection. Ces courbes sont obtenues avec l'utilisation de l'augmentation des données avec un modèle initialisé avec les poids appris sur ImageNet.

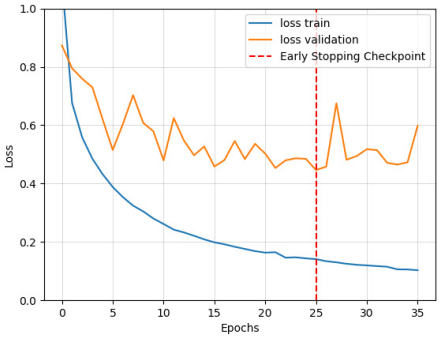
Tableau 5.3 – Résultats quantitatifs obtenus de la classification des parcelles (prairies, vignes, vergers traditionnels et vergers intensifs) avec l'approche globale $G - STR$ (Taux de classification – TC, écart-type – ET).

		initialisation aléatoire		*Fine tuning*	
Représentation		**TC**	**ET**	**TC**	**ET**
$\Re_{snake}$	sans AD	**70.00**	**1.70**	79.94	2.06
$\Re_{spiral}$	sans AD	68.92	2.50	77.23	1.42
$\Re_{Hilbert}$	sans AD	69.23	2.82	**81.69**	**1.88**
$\Re_{snake}$	avec AD	**81.12**	**2.37**	91.43	1.58
$\Re_{spiral}$	avec AD	76.05	2.53	89.43	1.61
$\Re_{Hilbert}$	avec AD	80.51	2.28	**91.69**	**0.91**

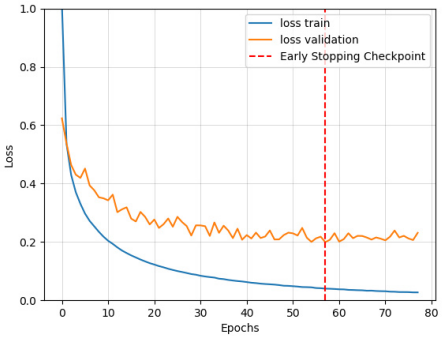
Nous rajoutons une stratégie naïve pour créer les STR qui sera considérée comme méthode de base. Cette méthode consiste à créer des STR en sélectionnant les pixels du support $\mathcal{D}$ de la STI de façon aléatoire, notée $Rand$. Nous fixons le nombre de pixels par STR à $L = 10$ tout en faisant varier le nombre de représentations comme dans l'approche $MS - STR$. Une telle méthode peut permettre d'interpréter la STR comme un sac de pixels temporels.

La figure 5.19 présente les courbes de perte quand le modèle est entraîné sur les STR de l'approche globale $G - STR$ avec l'augmentation des données. Notons que le modèle est à chaque fois initialisé avec les poids appris sur ImageNet. Nous remarquons que les trois courbes sont similaires et qu'elles ne dépassent pas les 20 *epochs* d'entraînement. La validation est aussi au même niveau dans les trois courbes. Cela justifie que les différentes courbes remplissant l'espace sont quasiment équivalentes en terme de résultats quantitatifs.

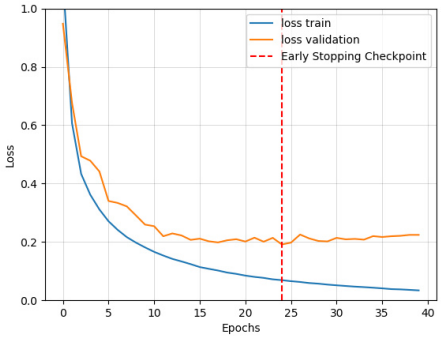
Le tableau 5.3 résume les scores obtenus avec la méthode globale $G - STR$. Cette approche semble être moins efficace que l'approche locale $MS - STR$. Cela peut être dû en partie à la petite taille de l'ensemble d'apprentissage disponible. Avec la technique d'aug-



(a) $Rand$



(b) $RW(50)_{70\%}$



(c) $RW(100)_{70\%}$

Figure 5.20 – Courbes de pertes obtenues quand le modèle est entraîné sur les STR de $Rand$, $RW(50)$ et $RW(100)$. Pour l'application de télédétection, le modèle est initialisé avec les poids appris sur ImageNet.

Tableau 5.4 – Résultats quantitatifs obtenus de la classification au niveau STR (prairies, vignes, vergers traditionnels et vergers intensifs) avec l'approche locale $MS-STR$ (Taux de classification – TC, écart-type – ET).

	N_{seg}	Initialisation aléatoire		*Fine tuning*	
Représentation	**Entraîn. / Test**	**TC**	**ET**	**TC**	**ET**
$Rand$	10%	82.27	2.53	86.33	1.21
	20%	85.57	2.98	87.23	1.97
	50%	**87.90**	**0.48**	**89.04**	**1.12**
	70%	86.12	1.68	88.66	1.04
$RW(10)$	10%	77.29	0.76	84.10	1.15
	20%	80.41	0.97	87.13	0.73
	50%	83.13	1.42	88.25	0.58
	70%	**85.64**	**0.54**	**89.23**	**0.83**
$RW(50)$	10%	80.48	2.85	88.47	0.58
	20%	83.46	0.62	89.05	1.06
	50%	87.02	1.69	89.84	1.02
	70%	**89.07**	**1.23**	**90.12**	**0.95**
$RW(100)$	10%	83.44	0.93	89.82	0.87
	20%	86.62	0.98	90.56	1.01
	50%	87.78	1.60	89.83	1.18
	70%	**88.25**	**1.69**	**90.35**	**1.50**

mentation des données, les scores s'améliorent mais restent en dessous de ceux obtenus avec $MS-STR$. Par ailleurs, les STR de l'approche $G-STR$ sont basées sur une topologie spatiale 4-connexité par rapport à $MS-STR$ qui est basée sur la 8-connexité. Cela peut expliquer en partie les faibles résultats observés. Enfin, le caractère aléatoire de l'orientation des RW permet d'avoir une vue isotrope de la région locale dans l'approche $MS-STR$. Le *fine tuning* à son tour fait augmenter tous les scores avec ou sans augmentation de données.

La figure 5.20 illustre les courbes de perte obtenues quand le modèle est entraîné sur les STR générées avec $Rand$, $RW(50)$ et $RW(100)$. Notons que le modèle est à chaque fois initialisé avec les poids appris sur ImageNet. La première remarque à faire est que la courbe obtenue avec la stratégie $Rand$ contient beaucoup d'oscillations avec une erreur de validation plus élevée que les autres. Les courbes des $RW(50)$ et $RW(100)$ sont meilleures que la précédente avec $RW(50)$ qui nécessite plus d'*epochs*. Nous remarquons que le modèle $RW(100)$ commence à sur-apprendre après 30 *epochs*.

Le tableau 5.4 résume les scores obtenus de la classification des STR en quatre classes : prairies ; vignes ; vergers traditionnels et vergers intensifs, avec la méthode $MS-STR$. À première vue, la méthode $Rand$ semble la plus mauvaise de toutes et ce même en aug-

Tableau 5.5 – Résultats quantitatifs obtenus de la classification au niveau parcelle (prairies, vignes, vergers traditionnels et vergers intensifs) avec l'approche locale $MS-STR$ (Taux de classification – TC, écart-type – ET).

	N_{seg}	Initialisation aléatoire		*Fine tuning*	
Représentation	**Entraîn. / Test**	**TC**	**ET**	**TC**	**ET**
$Rand$	10%	81.84	2.26	88.87	1.56
	20%	85.33	3.68	89.02	1.38
	50%	**88.71**	**0.87**	**90.66**	**0.85**
	70%	88.15	1.96	90.00	1.13
$RW(10)$	10%	80.30	1.63	90.51	0.48
	20%	84.61	1.58	91.48	0.75
	50%	87.23	2.61	92.56	0.95
	70%	**89.28**	**0.96**	**93.07**	**1.02**
$RW(50)$	10%	81.64	3.31	91.07	2.53
	20%	84.82	1.32	93.80	1.57
	50%	89.33	0.92	94.06	1.44
	70%	**90.71**	**1.05**	**94.80**	**1.57**
$RW(100)$	10%	83.89	0.80	92.50	1.05
	20%	88.71	1.27	93.20	0.65
	50%	89.12	1.86	94.21	1.19
	70%	**89.53**	**2.10**	**94.64**	**0.80**

mentant le nombre de représentations. Le *fine tuning* apporte une légère amélioration mais la méthode $Rand$ reste une méthode naïve. Par contre quand nous utilisons les RW, les scores s'améliorent et dépassent ceux de la méthode $Rand$. Ceci montre l'intérêt de la prise en compte de l'information spatiale avec la méthode proposée. Quand le modèle est entraîné avec des poids initialisés aléatoirement, nous remarquons que les scores augmentent à chaque fois que le nombre de représentations augmente. En revanche, si nous considérons des segments très longs avec beaucoup de représentations, le modèle a tendance à sur-apprendre au fur et à mesure sur les données. Par ailleurs, si le point de départ de l'entraînement se fait à partir des poids appris de ImageNet, tous les scores s'améliorent en atteignant 90.35% avec $RW(100)$.

Le tableau 5.5 résume les scores obtenus de la classification au niveau parcelle avec la méthode $MS-STR$. L'obtention de ces scores est réalisée en appliquant la stratégie de décision expliquée dans la section 5.4. Pour cela, nous moyennons les probabilités retournées par le modèle de toutes les STR associées à chaque parcelle. Les scores de la méthode $Rand$ ont augmenté d'au moins 1%. Par contre avec les RW, une augmentation moyenne de 3% pour tous les scores et ce avec / sans *fine tunning*.

Nous observons ainsi que le CNN fournit de meilleurs scores quand il est initialisé avec les poids de ImageNet. La différence est d'environ 4% pour le $MS-STR$ et d'environ 10% pour le $G-STR$. Une telle différence pour la $G-STR$ peut être due à la plus faible quantité de données. En outre, l'apprentissage est beaucoup plus rapide lorsque le *fine tuning* est utilisé. Cela confirme l'intérêt d'avoir une bonne initialisation des poids d'un CNN.

En guise d'étude comparative, nous allons par la suite appliquer des méthodes de l'état-de-l'art sur les mêmes données afin de pouvoir classer la méthode proposée par rapport à elles. Les méthodes sélectionnées sont :

- $TempCNN$ [98] dédié à la classification des pixels temporels où des convolutions $1D$ sont appliquées seulement sur le domaine temporel ;
- baML, une méthode hybride qui traite d'un coté le domaine temporel avec des convolutions $1D$ et d'un autre coté le domaine spatial en incluant des caractéristiques spatiales. Cette dernière est proposée dans [84] ;
- un RNN composé de 3 couches de $LSTM$ avec 256 états cachés et une couche entièrement connectée, inspiré de [59] ;
- un RNN convolutif composé de 2 couches de $ConvLSTM$ avec 64 états cachés et d'une couche de convolution suivie d'un ReLu et d'une normalisation par lots, inspiré de [111]. Le classificateur est entièrement convolutif suivi d'une mise en commun de la moyenne globale ;
- $3D$-SqueezeNet [71], est une extension $3D$ du réseau SqueezeNet, proposée pour la reconnaissance d'actions humaines dans les vidéos. Dans son implémentation [71], l'entrée est une série de 16 images. Ici, nous sélectionnons 16 images en utilisant un pas régulier. Pour des raisons liées à la quantité des données, nous initialisons le modèle avec les poids appris sur les données Jester [3].

Nous avons employé le même protocole de validation avec les méthodes de l'état-de-l'art (comme expliqué dans la section 5.6.2). Le tableau 5.6 résume les scores obtenus avec les méthodes de l'état-de-l'art sélectionnées. Les meilleurs scores sont obtenus avec $TempCNN$ [98] et baML [84]. $TempCNN$ [98] a été adapté à la longueur des séquences temporelles qui sont égales à 224. La taille des filtres utilisés est de 11. Concernant baML [84], l'inclusion de l'information a porté ses fruits sachant que le CNN temporel utilisé est très léger par rapport à $TempCNN$. Le résultat du $LSTM$ [59] est en troisième position malgré leur réputation sur la capacité de mémoire longue. $3D$-SqueezeNet [71] donne des résultats très décevants. Cela peut venir du fait que, les CNN $3D$ nécessitent de grandes bases de données afin de pouvoir les entraîner efficacement. De plus, nous rappelons que la majorité des parcelles ont une petite taille, comme indiqué dans le tableau 3.1. $ConvLSTM$ [111] est en dernière position du classement.

En plus du taux de reconnaissance de classification, le temps d'inférence doit également être considéré. Pour cela, le tableau 5.7 résume les temps d'inférence pour toutes les mé-

3. La base Jester est une collection de vidéos des gestes de la main capté à partir d'ordinateurs portables (https://20bn.com/datasets/jester)

Tableau 5.6 – Résultats quantitatifs obtenus avec les méthodes de l'état-de-l'art et notre meilleure méthode pour l'application de télédétection. Les résultats sont triés dans l'ordre décroissant (Taux de classification – TC, écart-type – ET).

Méthodes	**TC**	**ET**
$MS-STR$ $RW(50)_{70\%}$	**94.80**	**1.57**
$TempCNN$ [98]	**92.98**	**0.89**
$G-STR$ $\Re_{Hilbert}$	91.69	0.91
baML [84]	91.25	0.53
$3D$-SqueezeNet [71]	85.33	1.19
$LSTM$ [59]	83.48	2.29
$ConvLSTM$ [111]	74.66	1.56

Tableau 5.7 – Temps d'inférence en secondes pour les meilleurs modèles par rapport aux méthodes de l'état-de-l'art (classés par ordre croissant).

Méthodes	**Temps moyen (en secondes)**
G-STR $\Re_{Hilbert}$	2.72
baML [84]	11.91
$TempCNN$ [98]	13.50
MS-STR $RW(50)$	22.57
$ConvLSTM$ [111]	23.16
$3D$-SqueezeNet [71]	26.50
$LSTM$ [59]	26.96

thodes de l'état-de-l'art et nos deux meilleurs modèles. Le modèle le plus rapide est $\Re_{Hilbert}$ qui est celui de l'approche globale $G-STR$ car le nombre de STR par parcelle est limité. En deuxième et troisième positions viennent les méthodes baML [84] et $TempCNN$ [98]. Le modèle $RW(50)_{70\%}$ de l'approche locale $MS-STR$ arrive en quatrième position car le nombre des STR est plus élevé. Enfin, $ConvLSTM$ [111], $3D$-SqueezeNet [71] et $LSTM$ [59] sont en dernières positions en raison de la complexité et de la gestion de la mémoire de ces méthodes.

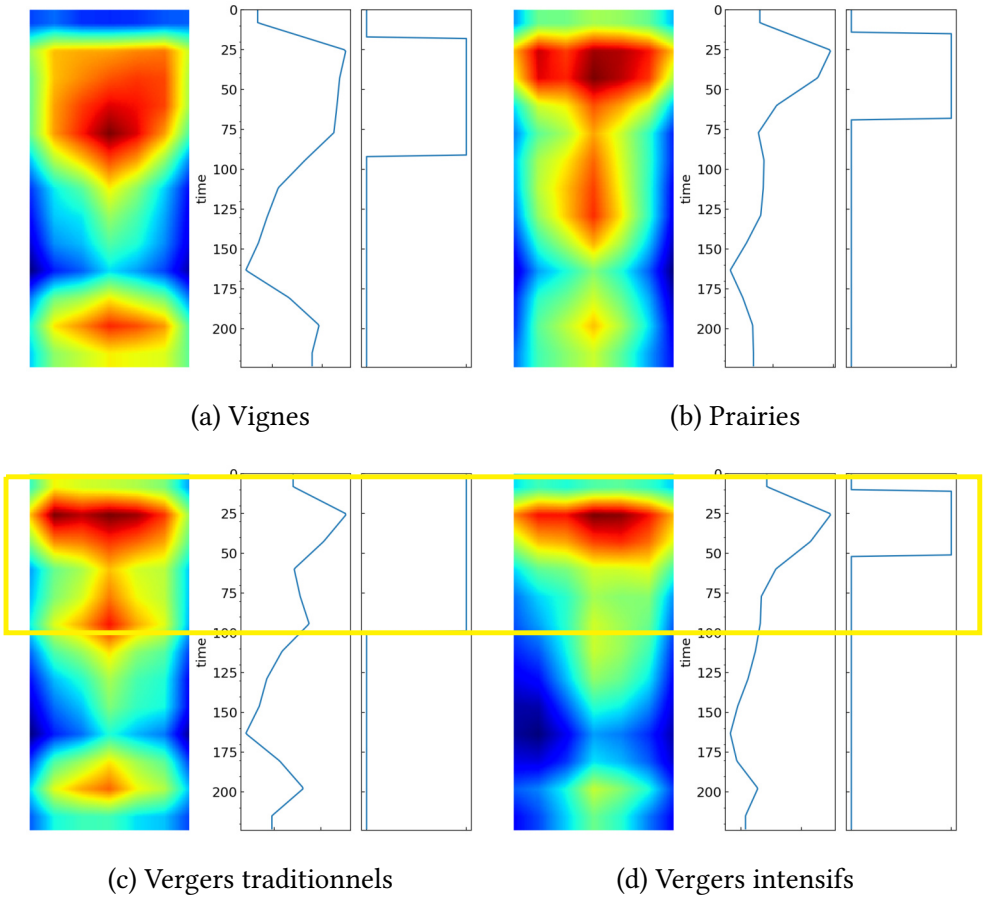
Afin de mieux cerner le problème des classes déséquilibrées comme présenté dans le tableau 5.1, une analyse des résultats par classe est menée et le tableau 5.8 présente le rappel, la précision et le F1-Score des méthodes de l'état de l'art et de nos meilleures méthodes. En se focalisant sur le F1-Score moyen, les deux meilleures méthodes sont $RW(50)_{70\%}$ de l'approche locale $MS-STR$ (en première position), suivie de $\Re_{Hilbert}$ de l'approche globale $G-STR$. Par ailleurs, les méthodes $TempCNN$ [98], baML [84], $LSTM$ [59] et

Tableau 5.8 – Résultats obtenus par classe pour l'application de télédétection (précision – P, rappel – R et F1-Score – F1).

Classes		Prairies	Vignes	Vergers trad.	Vergers int.	Moyenne
MS-STR - $RW(50)_{70\%}$	P	96.08	99.47	78.81	88.02	90.59
	R	95.80	98.58	91.42	81.01	91.70
	$F1$	**95.94**	99.02	**84.53**	**84.16**	**90.91**
G-STR - $\Re_{Hilbert}$	P	94.56	96.70	71.40	79.71	85.59
	R	92.66	96.84	85.18	75.89	87.64
	$F1$	93.59	96.75	77.16	77.72	86.30
$TempCNN$ [98]	P	90.98	99.82	89.47	85.42	91.42
	R	99.29	99.82	48.57	73.04	80.18
	$F1$	94.94	**99.82**	61.78	78.68	83.80
baML [84]	P	93.51	96.75	67.17	74.12	82.88
	R	97.14	99.12	54.07	62.56	78.22
	$F1$	95.28	97.92	58.86	67.72	79.94
$LSTM$ [59]	P	95.58	96.07	34.58	48.80	68.75
	R	82.76	98.77	47.40	67.69	74.15
	$F1$	88.70	97.40	39.91	56.55	70.64
$ConvLSTM$ [111]	P	84.95	84.94	12.07	26.88	52.21
	R	80.66	93.33	11.85	31.28	54.28
	$F1$	82.75	88.94	11.96	28.91	53.14
$3D$-SqueezeNet [71]	P	91.41	94.79	39.84	56.21	70.56
	R	88.09	98.77	29.63	96.74	78.30
	$F1$	89.58	96.73	33.96	62.36	70.65

$3D$-SqueezeNet [71] ont tendance à confondre les deux types de vergers qui ne sont pas les classes les plus représentées. Ces types de pratique de gestions agricoles sont caractérisés par une forte structuration spatiale des arbres et une évolution temporelle qui est souvent liée aux interventions agricoles de l'agriculteur.

D'après les scores du tableau 5.8, nous remarquons que toutes les approches ont tendance à confondre les prairies et les vergers. La méthode $MS-STR$ a tendance à classer les vergers traditionnels comme intensifs car le rappel est très élevé dans cette classe. Afin de mieux analyser ce comportement et aussi de comprendre ce choix de la part du CNN, nous utilisons les mécanismes d'attention proposées précédemment pour déterminer quelles parties des STR sont les plus discriminantes en fonction des classes. Cela peut également permettre d'identifier certaines erreurs.



(a) Vignes (b) Prairies

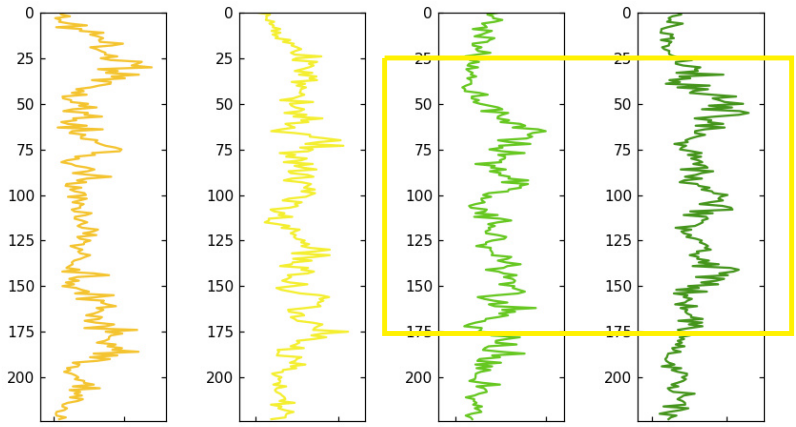
(c) Vergers traditionnels (d) Vergers intensifs

Figure 5.21 – Illustration de l'attention temporelle calculée avec le meilleur modèle de la méthode proposée ($RW(50)_{70\%}$ de l'approche $MS-STR$) pour l'application de télédétection. La moyenne des cartes d'attention des quatre classes est donnée avec leurs profils d'attention temporelle associés et leurs binarisations. Les rectangles jaunes représentent les plages temporelles d'intérêt considérées dans une étude à 2 classes (vergers traditionnels vs. intensifs)

Attention temporelle

Les figures 5.21 illustrent les cartes d'attention temporelle obtenues sur les quatre classes agricoles avec le modèle $RW(50)$ quand le nombre de STR est égal à 70% des pixels des parcelles, noté $RW(50)_{70\%}$ de l'approche locale $MS-STR$. La moyenne des cartes d'attention des quatre classes est fournie avec les profils d'attention temporelle associés et leurs binarisations.

Pour mieux évaluer l'intérêt de cette approche, nous nous concentrons ici sur une étude à 2 classes, impliquant des vergers traditionnels et intensifs. Parmi les profils d'attention temporelle associés, nous sélectionnons la période la plus discriminante. Cette période correspond à un intervalle temporel compris entre les temps 1 et 100 (*i.e.*, janvier à mi-juin).



(a) Vignes (b) Prairies (c) Vergers traditionnels (d) Vergers intensifs

Figure 5.22 – Illustration de l'attention temporelle pour les quatre classes obtenues avec $TempCNN$ [98] pour l'application de télédétection. Le rectangle jaune représente la plage temporelle d'intérêt considérée dans une étude à deux classes (vergers traditionnels vs. intensifs).

En appliquant le modèle et son entraînement sur cette nouvelle plage temporelle (comme il est présenté avec les rectangles jaunes dans la figure 5.21) les résultats s'améliorent. Plus précisément, l'erreur est réduite d'environ 7%. Cela montre l'intérêt de l'étude des cartes d'attention en écartant les périodes non significatives de l'année selon le problème considéré.

À titre de comparaison, nous avons mené une étude en générant des cartes d'attention temporelle sur les quatre classes agricoles avec $TempCNN$ [98]. La figure 5.22 illustre les résultats obtenus. Le rectangle jaune représente à nouveau la nouvelle plage temporelle (entre 25 et 195 (*i.e.*, février à novembre)) considérée dans l'étude à 2 classes. En réentraînant le modèle sur ces intervalles restreints, la diminution de l'erreur est d'environ 5% avec $TempCNN$ [98].

En outre, la précision est dans les deux cas améliorée. Pour conclure cette étude, l'attention temporelle permet de confirmer que la saison du printemps était plus significative que le reste de l'année pour effectuer la discrimination entre les deux différentes pratiques de gestion agricole considérées.

Attention spatiale

La figure 5.23 illustre les cartes sémantiques spatiales sur quelques exemples de paysages impliquant des prairies. Les prairies sont de grandes zones où les pratiques agricoles peuvent varier et évoluer dans le temps (reboisement, nouvelles cultures) en fonction des

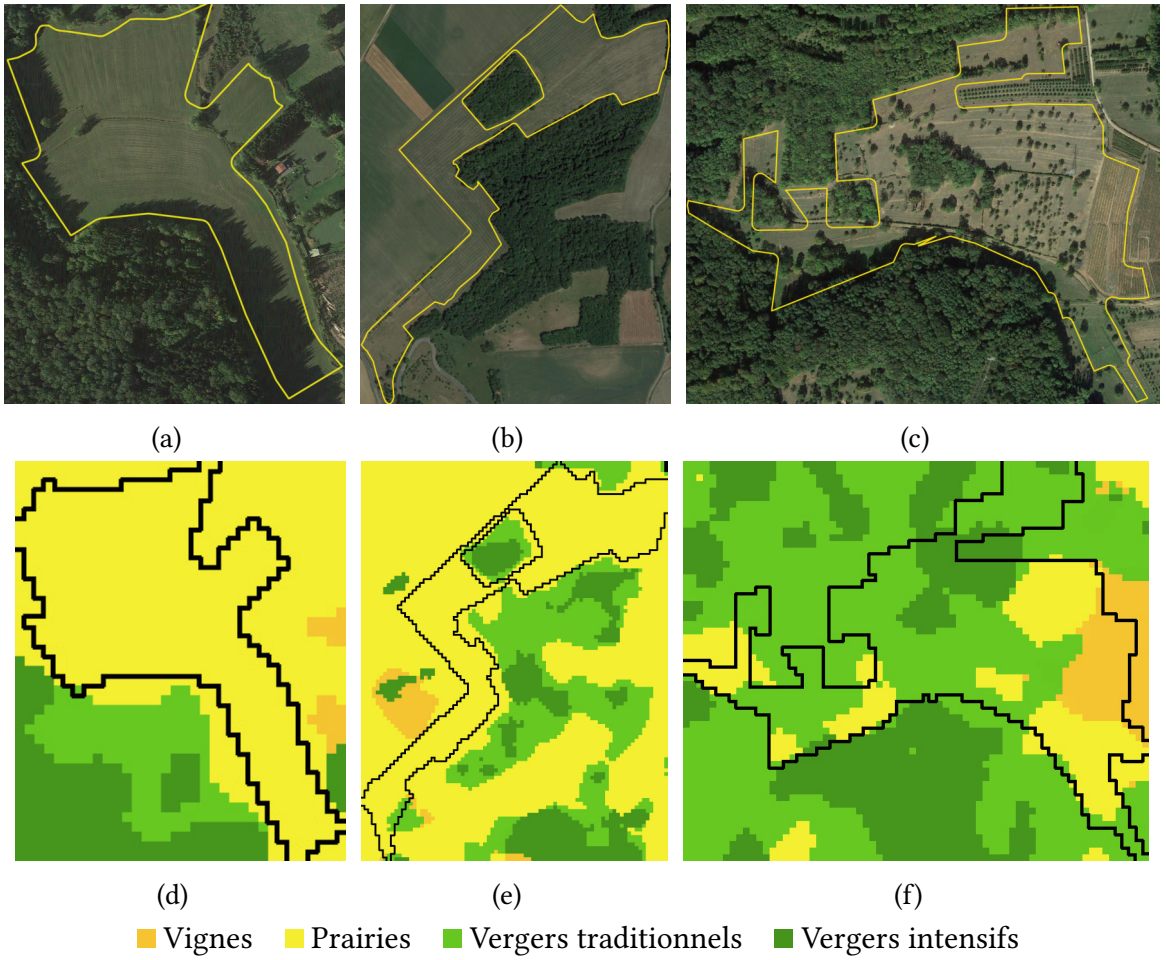


Figure 5.23 – Illustration des cartes de segmentation sémantique obtenues avec la méthode proposée : (a, b, c) Trois prairies représentées sur une image à très haute résolution spatiale provenant de Google Earth (les limites des prairies sont en jaune) ; (d, e, f) Cartes de segmentation sémantique basées sur l'attention spatiale avec notre meilleur modèle $MS-STR$ obtenu avec $RW(10)$.

besoins des propriétaires des parcelles. Elles constituent des objets d'intérêt très hétérogènes où une seule étiquette de classe peut ne pas être très cohérente et cette hétérogénéité peut conduire le classificateur à des erreurs. À des fins de visualisation, les cartes sémantiques sont comparées à des images à très haute résolution spatiale provenant de Google Earth, comme présentées dans la figure 5.23.(a, b, c). Cela permet d'observer les détails dans le voisinage spatial des parcelles. Nous avons choisi de réaliser les cartes sémantiques avec le meilleur modèle $RW(10)$ de l'approche $MS-STR$ car la longueur 10 du segment permet d'être précise en terme d'information spatiale enrichissant le pixel temporel. Autrement dit, limiter l'information spatiale au pixel et éviter d'avoir un mélange de culture dans la STR. Les cartes sémantiques obtenues sont illustrées sur la figure 5.23.(d, e, f). Le modèle arrive à bien classer les prairies (a, b) alors que la prairie (c) est mal classée en tant que verger traditionnel. En analysant visuellement les cartes sémantiques, presque tous les pixels des

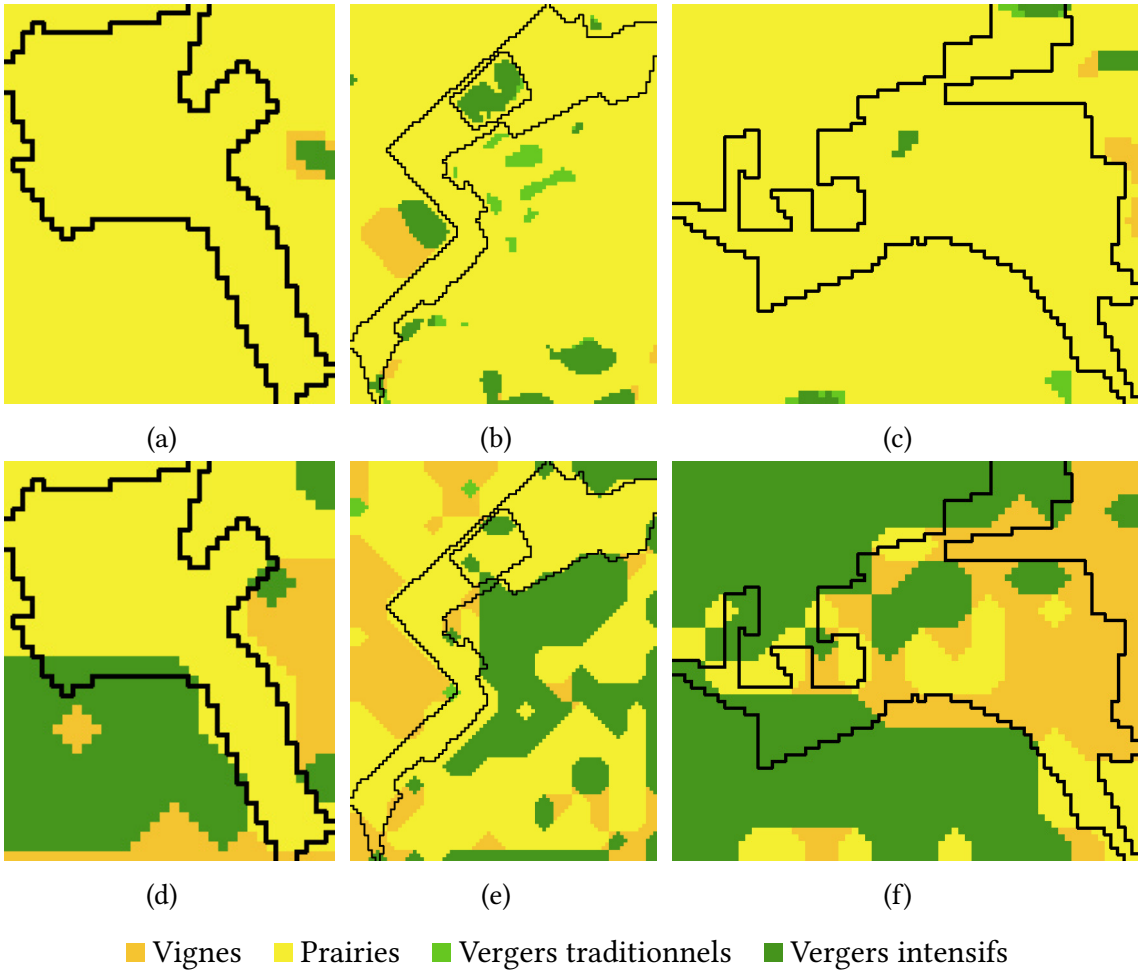


Figure 5.24 – Illustration des cartes de segmentation sémantique obtenues avec les méthodes de l'état-de-l'art : (a, b, c) Cartes de segmentation sémantique obtenues avec $TempCNN$ [98] ; (d, e, f) Cartes de segmentation sémantique spatiales obtenues avec $3D$-SqueezeNet [71].

prairies (a, b) sont étiquetés comme des prairies (couleur jaune), comme représenté dans (d) et (e). Afin de bien comprendre l'échec du modèle pour la prairie (c), nous remarquons sur l'image (c) que la prairie comprend de nombreux arbres isolés. Cela explique pourquoi sur la carte sémantique (f), une majorité de pixels sont étiquetés comme verger traditionnel (couleur vert clair). La partie droite de cette ROI contient également des vignes et elle est bien étiquetée grâce au mécanisme d'attention spatiale.

La figure 5.24 illustre les cartes de segmentation sémantique obtenues avec deux méthodes de l'état-de-l'art. La première méthode est $TempCNN$ [98] (a, b, c) et la deuxième est $3D$-SqueezeNet [71] (d, e, f). $TempCNN$ arrive à bien classer les trois prairies. Mais les cartes de segmentation obtenues avec ce modèle sont presque entièrement homogènes. Quelques pixels sont classés comme des vergers ou des vignes. Pour avoir une segmentation avec $3D$-SqueezeNet, nous avons procédé par une classification par patchs de taille

5×5. Ensuite, les patchs sont classés et nous affectons une couleur spécifique en fonction de l'étiquette prédite. $3D$-SqueezeNet arrive à bien classer les prairies de la figure 5.23.(a, b) mais se trompe dans la prairie de la figure 5.23.c. Toutefois, les pixels de forêt ont bien été étiquetés comme des vergers intensifs dans les cartes de segmentation sémantique. Un point commun entre toutes les cartes de segmentation est que dans la prairie de la figure 5.23.b, une même zone est toujours labellisée comme vigne. Au-delà de ce détail, la concordance entre les observations du terrain (à partir de l'image satellite Google Earth) et les résultats obtenus avec $MS-STR$ par rapport à l'état de l'art confirme à nouveau l'intérêt du mécanisme d'attention spatiale proposé. Cette expérience met en évidence la façon dont les cartes sémantiques peuvent expliquer le contexte spatial de la conclusion, permettant par exemple d'étudier les parcelles mal-classées ou hétérogènes.

Analyse des filtres convolutionnels appris par le CNN

Dans cette partie, nous appliquons la stratégie présentée dans la section 5.5.2 afin d'analyser les filtres convolutionnels appris par le CNN. En partant d'une initialisation des poids du modèle de façon aléatoire ou avec les poids d'ImageNet (où les filtres appris sont purement spatiaux), nous souhaitons étudier comment l'entraînement du réseau à partir des STR peut conduire à l'apprentissage de caractéristiques $2D$ capturant des informations spatio-temporelles. Nous avons choisi de n'utiliser que le meilleur modèle de $MS-STR$ qui est *RW(50)*$_{70\%}$. L'étude menée dans cette partie se limite à l'analyse des filtres de la première couche de convolution de SqueezeNet. Cette couche est constituée de 64 filtres qui sont présentés sur la figure 5.25. Nous utilisons par la suite les images synthétiques I_s et I_t, présentées dans la section 5.5.2, pour analyser la nature des filtres.

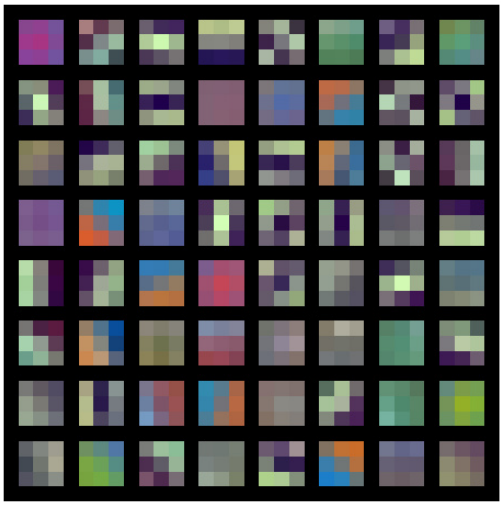
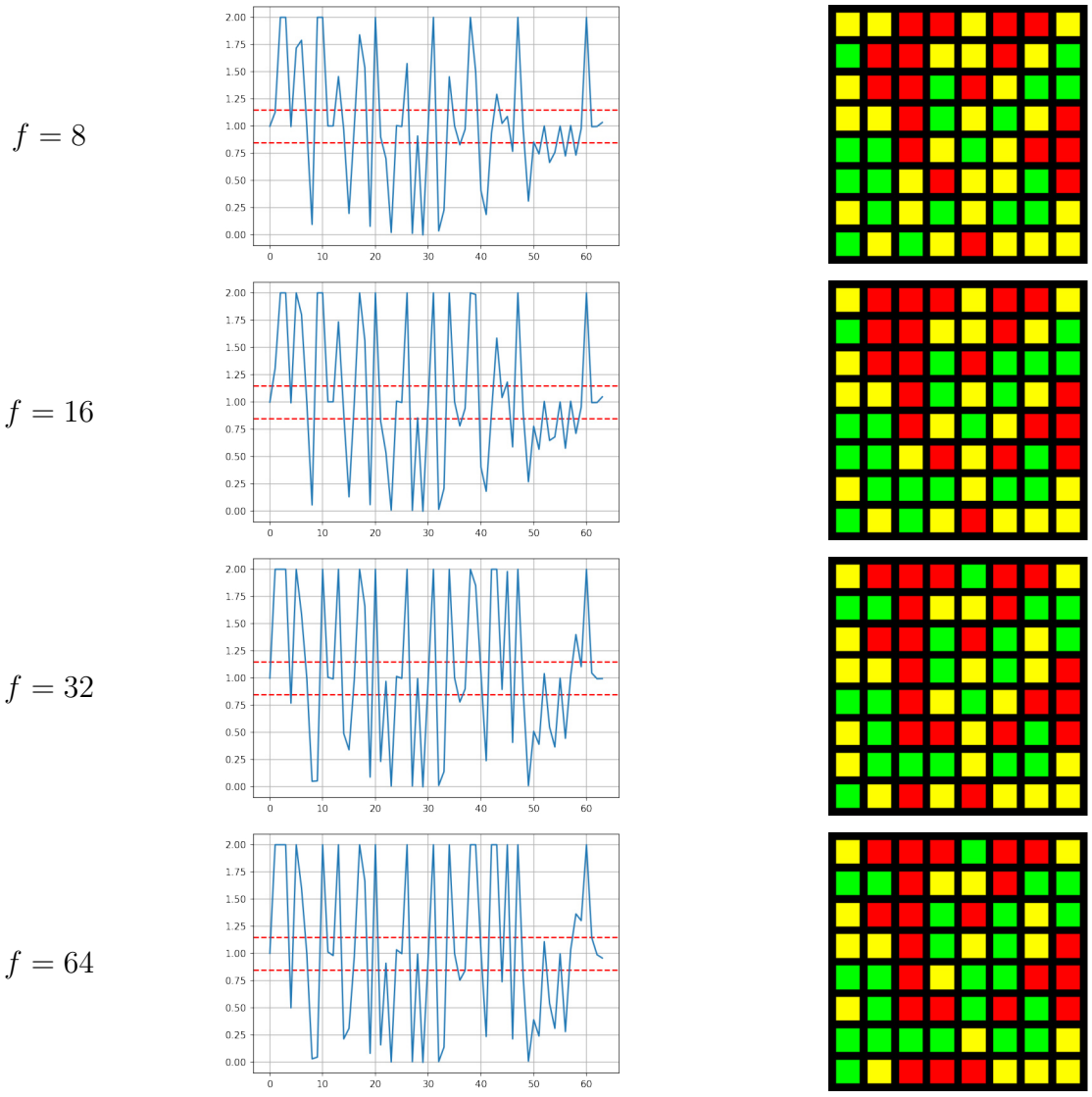


Figure 5.25 – Visualisation des 64 filtres de la première couche de convolution de SqueezeNet du meilleur modèle de $MS-STR$ dans l'application de télédétection.

Les images synthétiques sont ensuite données au CNN et nous récupérons les caractéristiques de convolution de la première couche. Par la suite, l'énergie est calculée pour chacune des 64 cartes de caractéristiques et ce pour I_s et I_t. En appliquant la formule 5.12, le rapport spatio-temporel R_{st} des deux énergies est obtenu par lequel la nature des filtres est identifiée. La figure 5.26.a présente les résultats des différents ratios R_{st} associés aux différentes valeurs de f. L'identification de la nature des filtres (spatial, temporel ou spatio-

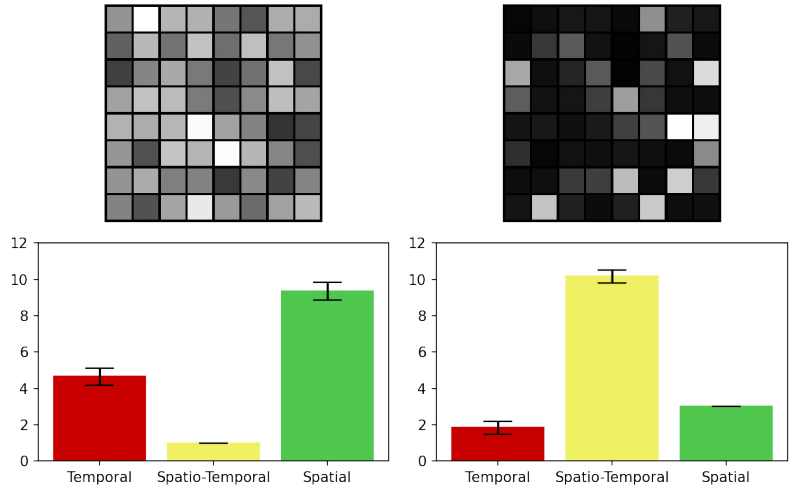


(a) Les ratios de l'énergie R_{st} obtenus pour les différentes valeurs de f et les différents filtres.

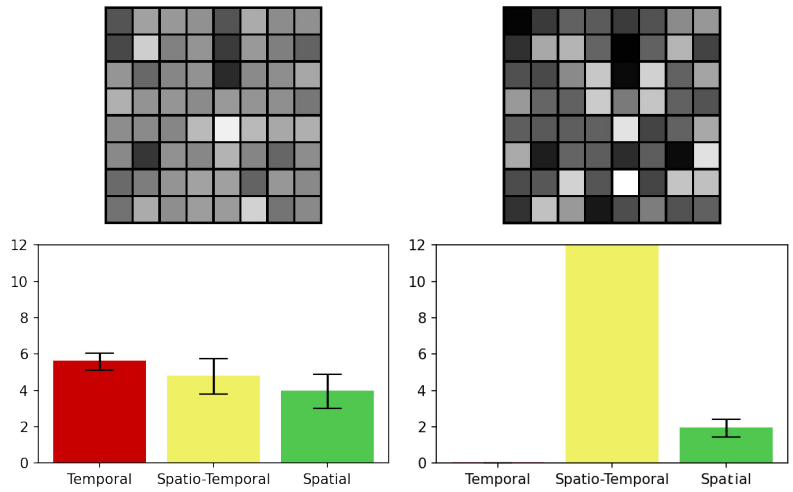
(b) Classification des 64 filtres pour les différentes valeurs de f.

Figure 5.26 – Illustration des ratios des 64 énergies associées aux caractéristiques obtenues avec la première couche de convolution du CNN et leur classification selon la nature de l'information (les résultats sont triés selon la valeur de f – de 8 à 64).

temporel) est faite en fonction du ratio R_{st}. Cela est réalisé en fixant les valeurs μ et ν. Dans notre cas, nous avons choisi de les fixer tous les deux à la valeur 0.15 afin d'avoir une équivalence de sélection des filtres spatio-temporels. La classification des filtres est réalisée selon les conditions présentées dans la section 5.5.2. Nous avons affecté une couleur pour chaque nature. Les couleurs rouge, vert et jaune représentent respectivement les natures temporelle, spatiale et spatio-temporelle. La figure 5.26.b illustre les résultats de classification obtenus associés à chaque valeur f. En outre, nous observons que les filtres appris sont en équivalence quand f est égale à 8, 16 et 32. Mais nous observons qu'à chaque fois où f augmente, le nombre de filtres spatio-temporels diminue. Cela montre que les filtres s'adaptent au contenu des images.



(a) Verger intensif (b) Verger traditionnel



(c) Prairie (d) Vigne

Figure 5.27 – Les filtres les plus actifs selon les énergies calculées pour la classification de parcelles agricoles ; (haut) Valeurs d'énergies de chaque filtre ; (bas) La nature des filtres les plus actifs.

Nous analysons dans la suite quels sont les filtres les plus actifs pour quelques images associées à quatre parcelles agricoles : un verger traditionnel, un verger intensif, une prairie et une vigne. Pour ce faire, nous utilisons la classification des filtres quand f est égale à 8. Le but maintenant est d'identifier la nature des filtres les plus actifs lors de l'analyse de ces quatre types de parcelles. Pour ce faire, chaque STR est traitée individuellement en identifiant les filtres les plus actifs. Cela est réalisé en comptant la fréquence d'apparition de chaque nature parmi les 64, présentée sous la forme d'histogrammes. Nous avons limité le comptage aux 15 énergies les plus hautes afin de ne considérer que les filtres les plus significatifs. Afin d'avoir un résultat par parcelle, la méthode est appliquée pour chaque STR de la parcelle, ensuite la moyenne des histogrammes est calculée.

Les résultats obtenus sont présentés dans la figure 5.27. Les histogrammes associés aux vergers et à la vigne montrent bien que l'information spatiale est plus utilisée que l'information temporelle. Cela est visible par la forte fréquence des filtres spatiaux ou spatio-temporels. L'alignement des rangées d'arbres dans les vergers intensifs est matérialisée par l'utilisation plus importante des filtres spatiaux. Quant à la prairie, cette dernière se caractérise par l'homogénéité du comportement spatial tout au long de l'année, ce qui est confirmé par la faible utilisation des filtres spatiaux par rapport à l'importance des filtres temporels. Cette analyse montre que les informations temporelles et spatiales sont très importantes dans le problème de classification que nous considérons, justifiant l'utilisation de la méthode $Deep - STaR$ ainsi que les améliorations par rapport aux approches temporelles.

5.6.3.2 Application à la détection de violence à partir de vidéos

Nous passons maintenant à l'application de la méthode sur notre deuxième cadre applicatif lié à la classification des vidéos (violentes ou non violentes). Tout d'abord, une étude préliminaire est faite afin de trouver la bonne configuration des STR. Pour cela, nous avons utilisé différentes longueurs L qui sont 50, 100 et 224. Nous rappelons que nous avons fait varier le nombre de STR de test N_{seg}^{test} par vidéo à 10, 30, 50 et 100. Cette étude a été réalisée seulement avec la base RWF2000. La figure 5.28 présente les résultats obtenus par cette

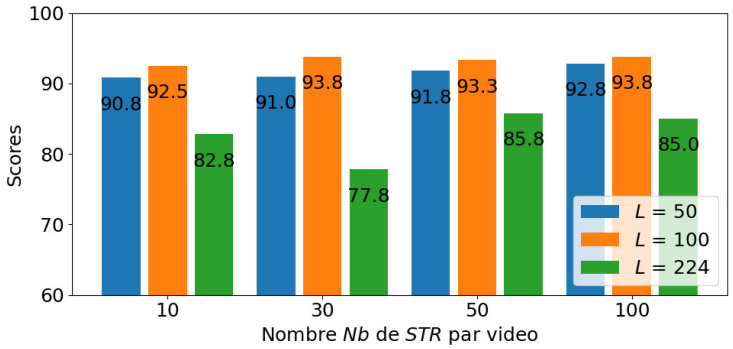


Figure 5.28 – Résultat de classification de l'étude préliminaire sur la base RWF2000.

étude. Les STR avec $L = 224$ donnent les résultats les plus faibles puis viennent les STR de longueur $L = 50$ et les meilleurs scores sont obtenus avec les STR de $L = 100$. Cela démontre l'importance du paramètre L et son impact sur les résultats. Si le segment est trop long ou trop court, le modèle n'arrive pas à s'adapter pour extraire de bonnes caractéristiques. Dans la suite, nous fixons la longueur $L = 100$ pour tester sur les autres jeux de données.

En plus de la méthode $MS - STR$ proposée, d'autres approches de l'état-de-l'art ont été appliquées à la même tâche comme étude comparative. Parmi les méthodes de l'état-de-l'art, nous nous sommes comparés à deux familles. La première famille concerne les méthodes basées sur les $CNN\ 3D$ qui sont :

- *Temporal Segment Networks* [139]. Ce dernier procède en divisant la vidéo en plusieurs clips ;
- $I3D$ [18] est un modèle qui n'utilise que la vidéo brute sans autres informations ;
- *Representation flow* [140] : ce modèle apprend à extraire des caractéristiques du flux optique ;
- *Flow Gated Network* [22] utilise la vidéo brute et son flux optique ;
- *Efficient Convolutional Network*, notée ECO [153] : ce dernier traite la vidéo image par image avec un $CNN\ 2D$. Puis les caractéristiques extraites pour toutes les images sont agrégées et classifiées avec un $CNN\ 3D$.

La deuxième famille de méthodes traite des nuages de points. Les méthodes sélectionnées sont :

- $PointNet++$ [102] : ce dernier traite les points groupe par groupe en appliquant *PointNet* sur chacun des groupes ;
- $PointConv$ [148] est un modèle qui utilise des convolutions $3D$ adaptées aux nuages de points ;
- *Dynamic Graph CNN* [141], notée $DGCNN$. Les convolutions de ce modèle sont appliquées sur les poids des arêtes des points liés ;
- *Skeleton Points Interaction Learning*, notée $SPIL$ [128] : cette méthode considère les squelettes comme des graphes et la convolution entre les points permet d'apprendre l'interaction entre ces graphes.

Le tableau 5.9 résume les résultats obtenus avec notre méthode et avec les méthodes de l'état-de-l'art. $MS - STR$ arrive à atteindre un taux de reconnaissance de 90% et ce avec tous les jeux de données considérés. En comparant avec les résultats des méthodes de l'état-de-l'art, $MS-STR$ est toujours classée entre la première et la troisième position par rapport aux méthodes concurrentes. Les résultats obtenus sur les jeux de données *Movies Fights* et *RWF2000* sont meilleurs que les autres car les vidéos dans ces ensembles ont toutes le même échantillonnage temporel. En revanche avec les jeux de données *Hockey Fights* et *Crowd Violence*, nos résultats sont surpassés par les autres méthodes. Cela est dû sans doute aux difficultés liées à chacun des jeux de données. Dans *Crowd Violence*, les vidéos n'ont pas la même durée et elles contiennent beaucoup de personnes, ce qui augmente la complexité

Tableau 5.9 – Résultats obtenus sur tous les jeux de données dans le cadre de l'application de reconnaissance de la violence avec notre méthode et celles de l'état-de-l'art.

Méthodes	***RWF2000***	***Movies fights***	***Hockey fights***	***Crowd Violence***
Méthode proposée $MS-STR-RW(100)$				
N_{seg}^{test}	ens. de test		valid. croisée	
10	92.5	**99.5**	92.2	88.2
30	**93.8**	96.0	93.9	90.6
50	93.3	98.5	93.6	89.0
100	**93.8**	98.5	94.4	89.8
CNN $3D$				
Temporal Segment Networks [139]	81.5	94.2	91.5	81.5
$I3D$ [18]	83.4	95.8	93.4	83.4
Representation flow [140]	85.3	**97.3**	92.5	**85.9**
Flow Gated Network [22]	**87.3**	n/a	**98.0**	88.8
ECO [153]	83.7	96.3	94.0	84.7
Nuage de points				
$PointNet++$ [102]	78.2	89.2	89.7	89.2
$PointConv$ [148]	76.8	91.3	89.7	89.2
$DGCNN$ [141]	80.6	92.6	90.2	87.4
$SPIL$ [128]	**89.3**	**98.5**	**96.8**	**94.5**

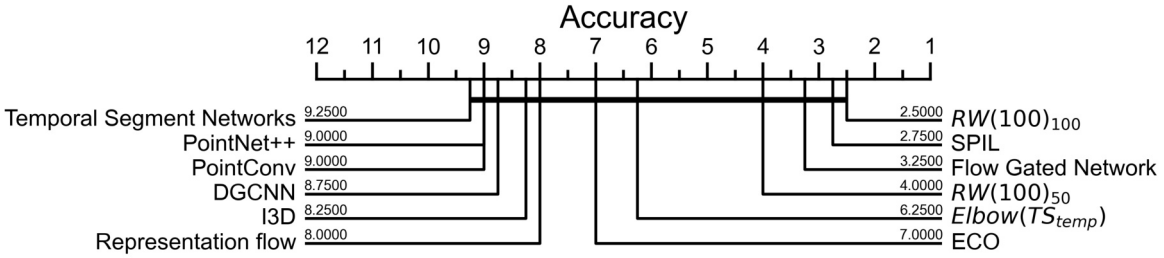
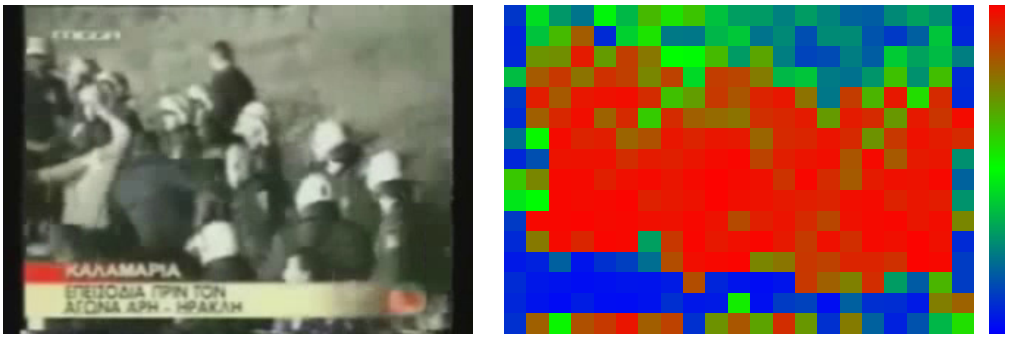


Figure 5.29 – Diagramme de différence critique (résultats de tous les ensembles de données) pour l'application de reconnaissance de la violence.

de la scène. Dans *Hockey Fights*, le mouvement de la caméra n'est pas stable et est parfois rapide. En cas de non-violence simulant le mouvement en arrière-plan, le mouvement de la camera est lent lorsque des événements violents se produisent. En outre, les mises à l'échelle de la scène peuvent être très différentes d'une vidéo à l'autre.

Pour comparer les méthodes d'une manière plus générale, nous utilisons un diagramme de différence critique [37] qui est illustré dans la figure 5.29. D'après ce diagramme, la méthode proposée fait partie des meilleures méthodes. La méthode proposée est en première position quand N_{seg}^{test} est égale à 100, notée $RW(100)_{100}$. En deuxième position, nous trouvons la méthode $SPIL$. Même en diminuant N_{seg}^{test} à 50 dans la prise de décision, notée $RW(100)_{50}$ sur le diagramme, notre approche se classe toujours bien. En guise de comparaison, nous ajoutons à ce diagramme la méthode de la stabilité temporelle où nous avons choisi les scores de TS_{temp} obtenus avec la loi du coude (section 4.6.2.2). TS_{temp} se classe aussi parmi les meilleures méthodes, elle se trouve juste dernière $RW(100)_{50}$. En conclusion, les méthodes sont toutes de qualité équivalente mais les résultats sont différents selon le jeu de données considéré.

En générant les STR de façon dense sur tout le domaine spatial $\mathcal{D}$ de la vidéo, une segmentation sémantique globale mettant en évidence les régions violentes peut également être obtenue. Pour ce faire, le domaine spatial $\mathcal{D}$ est divisé en petits patchs de taille w dans lesquels une STR est générée. Pour réaliser la segmentation, nous affectons les probabilités obtenues par le CNN aux pixels de chaque patch. La figure 5.30 illustre un exemple de la carte sémantique obtenue pour une vidéo issue de la base *Crowd Violence*. L'échelle de couleurs va du rouge, indiquant les zones les plus violentes (la probabilité de violence pour qu'une STR soit haute) au bleu (la probabilité de la non-violence pour qu'une STR soit haute).



(a) (b)

Figure 5.30 – Résultats de la classification sur une vidéo de foule issue de *Crowd Violence* : (a) une image de la vidéo ; (b) Carte de probabilité de la violence (échelle de couleurs : du rouge (violence élevée) au bleu (pas de violence)).

5.7 Bilan scientifique

Dans cette partie du manuscrit, nous avons proposé une méthode permettant de représenter des données $2D + t$ dans une ou plusieurs représentations planaires sous la forme d'images $2D$ qui sont utilisées par la suite pour l'apprentissage de caractéristiques et leurs implications dans un problème de classification des STI. Les représentations pla-

naires $2D$ sont construites tout en préservant partiellement les relations spatiales des pixels sans perdre l'information temporelle. Ensuite, des caractéristiques spatio-temporelles natives sont extraites par l'utilisation d'un CNN classique $2D$. L'utilisation d'un CNN $2D$ permet de bénéficier de poids pré-appris à partir d'autres bases plus grandes, comme ImageNet et leur ajustement avec des données spécifiques. Nous expliquons comment les utiliser pour classifier des ROI issues de STI. Deux stratégies sont proposées pour analyser les configurations spatiales préservées, des stratégies locales $MS-STR$ et globales $G-STR$ en fonction de l'objectif choisi.

La méthode a été expérimentée sur deux applications différentes. La première consiste en une application de télédétection dédiée à la classification de parcelles de cultures agricoles. Les scores obtenus sont meilleurs que ceux de l'état de l'art tout en mettant en évidence l'intérêt de la méthode proposée. L'intégration d'un mécanisme d'attention original permet d'adapter une plage temporelle plus discriminante pour les différentes classes thématiques. De plus, la génération de cartes sémantiques spatiales permet d'avoir une interprétation fine de la décision prise. La deuxième application consiste en la classification des vidéos de violence. La génération des STR a été adaptée afin que l'étiquetage des STR soit davantage précis et aussi pour que nous puissions traiter une masse de données raisonnable. Les premiers résultats sont encourageants et nous pensons qu'ils sont améliorables. Nous remarquons que la tâche de classification devient plus simple (en utilisant des CNN $2D$ au lieu de $3D$) et nous bénéficions également des modèles entraînés sur des jeux de données plus importants.

Les travaux présentés dans ce chapitre ont fait l'objet de plusieurs publications en conférences et en revues internationales :

- Mohamed Chelali, Camille Kurtz, et Nicole Vincent. **Violence detection from video under 2D spatio-temporal representations.** *International Conference on Image Processing*, 2021, pages 2593-2597.
- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Deep-STaR : Classification of image time series based on spatio-temporal representations**. *Computer Vision and Image Understanding*, 2021, pages 208–209 :103221.
- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Influence of data representations and deep architectures in image time series classification**. *International Journal of Pattern Recognition and Artificial Intelligence*, 2020, pages 2160001.
- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Classification of spatially enriched pixel time series with convolutional neural networks.** *International Conference on Pattern Recognition*, 2020, pages 5310–5317.

- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **From pixels to Random Walk based segments for image time series deep classification.** *International Conference on Pattern Recognition and Artificial Intelligence*, 2020, pages 339–351.
- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Image time series classification based on a planar spatio-temporal data representation.** *International Conference on Computer Vision Theory and Applications*, 2020, pages 276–283.

Conclusion et perspectives

Bilan et contributions

Les travaux réalisés durant cette thèse s'inscrivent dans le contexte de l'analyse de séquences temporelles d'images. En particulier des méthodes ont été développées dans le cadre de l'extraction de caractéristiques spatio-temporelles utilisées par la suite dans un problème de classification. Les contributions méthodologiques ont été testées et évaluées dans deux cadres applicatifs différents. Le premier consiste en l'analyse des séries temporelles d'images satellitaires et le deuxième concerne la classification de vidéos.

La première contribution de cette thèse repose sur l'analyse de la stabilité temporelle des séquences temporelles d'images. Pour cela, nous avons transformé la séquence temporelle d'images en une représentation moins volumineuse par laquelle trois caractéristiques différentes sont extraites. La transformation est basée sur un algorithme de compression nommé *Run Length Encoding*. Ce dernier compte le nombre de fois où une valeur est répétée successivement et a été appliqué au niveau du pixel temporel. Ensuite, différentes stratégies ont été étudiées pour que l'information spatiale soit prise en compte conduisant à des caractéristiques spatio-temporelles. La combinaison des trois caractéristiques en une image couleur a permis l'analyse des données avec une image de résumé au lieu d'analyser toutes les images individuellement.

La deuxième contribution méthodologique se focalise sur un changement de représentation des données. Les séquences temporelles d'images sont représentées en dimension $2D + t$. Nous proposons de transformer les données en une ou plusieurs représentations planaires prenant la forme d'images $2D$ qui contiennent les informations spatiales et temporelles. Différentes stratégies sont utilisées pour pouvoir garder des informations spatiales significatives. Grâce à ces nouvelles images, des caractéristiques spatio-temporelles sont ensuite extraites avec des réseaux de neurones convolutionnels $2D$. De plus, nous bénéficions des poids déjà appris sur de grandes bases d'images telle ImageNet. En plus de la représentation, deux stratégies d'analyse sont proposées qui ont pour but la compréhension des résultats obtenus par le réseau de neurones. La première stratégie concerne l'utilisation des cartes de saillance permettant de mettre en évidence la région spatiale de l'image la plus

utilisée par le CNN. Grâce à de telles cartes, deux mécanismes d'attention sont proposés. Le premier mécanisme analyse la plage temporelle la plus significative pour le réseau et le deuxième mécanisme permet de faire une segmentation sémantique avec un CNN conçu pour faire la classification. La deuxième stratégie a pour but d'analyser la nature des filtres (spatiale, temporelle ou spatio-temporelle). Cette dernière est basée sur l'utilisation d'images synthétiques et du calcul de l'énergie des réponses des différents filtres.

Les études expérimentales menées nous ont permis de comparer les méthodes proposées avec différentes méthodes de l'état-de-l'art en les appliquant sur les mêmes jeux de données et ce dans les deux cadres applicatifs. Les caractéristiques de stabilité combinées avec les pixels temporels ont permis d'avoir des scores meilleurs avec une forêt aléatoire qu'en utilisant un réseau de neurones convolutifs $1D$ dans le cas de l'analyse de la couverture urbaine. Ces dernières ont toutefois démontré leurs limites dans le cas de la classification de vidéos. Par contre, la deuxième contribution a dépassé toutes les méthodes de l'état-de-l'art lors de la classification des parcelles. Dans le cas de la classification des vidéos en violentes ou non violentes, l'utilisation du diagramme de différence critique a permis d'avoir un classement de toutes les méthodes avec la nôtre en première position.

Perspectives de recherche

Les travaux réalisés dans cette thèse concernent l'extraction de caractéristiques spatio-temporelles qui peuvent être *hand-crafted* ou apprises avec les méthodes d'apprentissage profond. Ils ont mis en évidence certaines perspectives de recherche.

Concernant l'analyse de la stabilité temporelle, nous prévoyons de poursuivre nos travaux sur la notion d'égalité utilisée pour décider si la valeur d'un pixel est stable dans le temps. Comme limite de notre travail, nous calculons actuellement notre représentation spatio-temporelle et les caractéristiques proposées à partir de données scalaires seulement, c'est-à-dire les pixels temporels caractérisés par le $NDVI$ ou la moyenne des niveaux de gris. Lorsque la caractéristique est vectorielle, par exemple toutes les bandes spectrales ou une combinaison d'indices de télédétection, il est possible de considérer de nouvelles définitions de l'égalité conduisant à une définition plus ou moins contraignante. La quantification des valeurs est aussi un problème à étudier car nous avons remarqué la difficulté d'interprétation du résumé avec les vidéos (contenu déformable ou avec continuité visuelle). De plus, il convient d'étudier l'impact des différents paramètres des relaxations proposées afin d'évaluer leurs capacités d'adaptation au contenu de la séquence temporelle d'images. L'utilisation des méthodes d'apprentissage nous permet de s'abstraire de trouver certains paramètres par une étude empirique comme la quantification des valeurs. Par contre, l'utilisation d'un réseau de neurones convolutif pourrait nous permettre de générer l'image du résumé. Cela pourrait se faire en utilisant un modèle qui prend en entrée la séquence d'images et génère l'image du résumé en sortie. De plus, l'image du résumé peut à son tour être utilisée comme entrée d'un modèle pour résoudre un problème de segmentation sémantique en utilisant une architecture de type $U - Net$. Dans le cadre de la classification de vidéos, nous nous proposons d'utiliser un modèle composé de deux têtes (qui comprend

deux entrées). La première entrée traiterait le résumé et la deuxième tête pourrait traiter soit la séquence temporelle d'images ou son flux optique de manière globale. Les caractéristiques extraites sont ensuite fusionnées et passées au classificateur du modèle.

La deuxième contribution consiste à représenter une séquence temporelle d'images par une ou plusieurs représentations $2D$. Ensuite, un $CNN\ 2D$ classique est utilisé pour apprendre des caractéristiques spatio-temporelles. Dans notre cas, nous n'avons utilisé que SqueezeNet dans nos expérimentations. Mais quel serait le résultat si un autre modèle était utilisé comme MobileNet ou VGG16 ? Dans le contexte de la détection de violence, la stratégie utilisée pour localiser les zones violentes dans les vidéos violentes pourrait être faite autrement. Par exemple, cette région est extraite par l'intersection du fort flux optique de toutes les images de la vidéo. Cela permet d'avoir une région plus restreinte que la stratégie utilisée. Une autre proposition est de rajouter la classe « arrière-plan » afin de pouvoir distinguer entre les régions violentes et non violentes. Du coté de la prise de décision de vidéo, un deuxième $CNN\ 2D$ peut être utilisé pour prédire le label. Par contre l'entrée du modèle ne serait pas les STR mais correspondrait à la segmentation sémantique globale qui met en évidence les régions violentes. Le mécanisme d'attention proposé n'a été appliqué que dans le cadre applicatif de télédétection. L'extension de celui-ci aux vidéos s'avère pertinent et surtout avec l'attention temporelle car elle permet de localiser temporellement la réalisation d'une action (violente ou autre dans un autre problème telle que la reconnaissance d'action). Quant à l'analyse des filtres, l'étude menée dans nos expérimentations n'ont concerné que les filtres de la première couche. Il serait également intéressant de faire cela au niveau des couches profondes du modèle.

Comme perspective générale, nos deux contributions méthodologiques peuvent être appliquées à d'autres problèmes. Par exemple, l'étude au fil du temps des modifications cellulaires dans le cas des images biomédicales. Une autre perspective est l'indexation des vidéos. Effectivement, pour la représentation compacte des vidéos, notre deuxième contribution méthodologique (basée sur les représentations $2D$) a l'avantage de traiter le domaine spatial de la vidéo de façon locale conduisant à différentes indexations.

Publications

Les résultats de nos travaux ont donné lieu à plusieurs publications scientifiques dans des revues internationales, en conférences internationales et nationales montrant l'intérêt de la définition des caractéristiques et ce par l'utilisation des deux méthodes.

Publications en revues internationales

- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Deep-STaR : Classification of image time series based on spatio-temporal representations**. *Computer Vision and Image Understanding*, 2021, pages 208–209 :103221.
- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Influence of data representations and deep architectures in image time series classification**. *International Journal of Pattern Recognition and Artificial Intelligence*, 2020, pages 2160001.

Publications en conférences internationales

- Mohamed Chelali, Camille Kurtz, et Nicole Vincent. **Violence detection from video under 2D spatio-temporal representations.** *International Conference on Image Processing*, 2021, pages 2593-2597.
- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Classification of spatially enriched pixel time series with convolutional neural networks.** *International Conference on Pattern Recognition*, 2020, pages 5310–5317.
- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **From pixels to Random Walk based segments for image time series deep classification.** *International Conference on Pattern Recognition and Artificial Intelligence*, 2020, pages 339–351.

- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Spatio-temporal stability analysis in Satellite Image Times Series.** *Second International Conference on Pattern Recognition and Artificial Intelligence.* 2020, pages 484–499.
- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Image time series classification based on a planar spatio-temporal data representation.** *International Conference on Computer Vision Theory and Applications*, 2020, pages 276–283.
- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Urban land cover analysis from satellite image time series based on temporal stability.** *Joint Urban Remote Sensing Event.* 2019, pages 1–4.

.

Publications en conférences nationales

- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Des pixels aux segments pour la classification de séries temporelles d'images via des réseaux de neurones convolutionnels.** *Congrès Reconnaissance des Formes, Image, Apprentissage et Perception* 2020.
- Mohamed Chelali, Camille Kurtz, Anne Puissant, et Nicole Vincent. **Classification de séries d'images via une représentation spatio-temporelle.** *Atelier sur l'Apprentissage Profond dans le cadre de la Conférence Extraction et Gestion des Connaissances*, 2020.

Bibliographie

[1] Radhakrishna Achanta et al. “SLIC Superpixels Compared to State-of-the-Art Superpixel Methods”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pages 2274-2282.

[2] Maryam N. Al-Berry et al. “Action Classification Using Weighted Directional Wavelet LBP Histograms”. In : *Advanced Intelligent System and Informatics*. Sous la direction de Tarek Gaber et al. Tome 407. Advances in Intelligent Systems and Computing. Springer, 2015, pages 15-24.

[3] Samaneh Aminikhanghahi et Diane J. Cook. “A survey of methods for time series change point detection”. *Knowledge and Information Systems* 51.2 (2017), pages 339-367.

[4] L. Andres, W.A. Salas et D. Skole. “Fourier analysis of multi-temporal AVHRR data applied to a land cover classification”. *Remote Sensing* 15.5 (1994), pages 1115-1121.

[5] Erchan Aptoula, Nicolas Courty et Sébastien Lefèvre. “An end-member based ordering relation for the morphological description of hyperspectral images”. In : *IEEE International Conference on Image Processing*. IEEE, 2014, pages 5097-5101.

[6] Gowtham Atluri, Anuj Karpatne et Vipin Kumar. “Spatio-Temporal Data Mining: A Survey of Problems and Methods”. *ACM Computing Surveys* 51.4 (2018), 83:1-83:41.

[7] Evert Attema et al. “The European GMES Sentinel-1 Radar Mission”. In : *IEEE International Geoscience & Remote Sensing Symposium*. IEEE, 2008, pages 94-97.

[8] Nicolas Audebert, Bertrand Le Saux et Sébastien Lefèvre. “How useful is region-based classification of remote sensing images in a deep learning framework?” In : *IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2016, pages 5091-5094.

[9] Lalit R. Bahl, Frederick Jelinek et Robert L. Mercer. "A Maximum Likelihood Approach to Continuous Speech Recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5.2 (1983), pages 179-190.

[10] Adeline Bailly et al. "Dense Bag-of-Temporal-SIFT-Words for Time Series Classification". In : *Advanced Analysis and Learning on Temporal Data - First ECML PKDD Workshop*. Sous la direction d'Ahlame Douzal Chouakria, José A. Vilar Fernández et Pierre-François Marteau. Tome 9785. Lecture Notes in Computer Science. Springer, 2015, pages 17-30.

[11] Tudor Barbu. "Pedestrian detection and tracking using temporal differencing and HOG features". *Computers and Electrical Engineering* 40.4 (2014), pages 1072-1079.

[12] Yoshua Bengio et Francois Gingras. "Recurrent Neural Networks for Missing or Asynchronous Data". In : *Advances in Neural Information Processing Systems*. Sous la direction de David S. Touretzky, Michael Mozer et Michael E. Hasselmo. MIT Press, 1995, pages 395-401.

[13] Simone Bianco et al. "Benchmark Analysis of Representative Deep Neural Network Architectures". *IEEE Access* 6 (2018), pages 64270-64277.

[14] M.J. Bishop et E.A. Thompson. "Maximum likelihood alignment of DNA sequences". *Journal of Molecular Biology* 190.2 (1986), pages 159-165.

[15] Lorenzo Bruzzone et Diego Fernández-Prieto. "Automatic analysis of the difference image for unsupervised change detection". *IEEE transactions on geoscience and remote sensing* 38.3 (2000), pages 1171-1182.

[16] Norbert Erich Buch, James Orwell et Sergio A. Velastin. "3D Extended Histogram of Oriented Gradients (3DHOG) for Classification of Road Users in Urban Scenes". In : *British Machine Vision Conference*. Sous la direction d'Andrea Cavallaro, Simon Prince et Daniel C. Alexander. British Machine Vision Association, 2009, pages 1-11.

[17] Arthur R. Butz. "Alternative Algorithm for Hilbert's Space-Filling Curve". *IEEE Transactions on Computers* 20.4 (1971), pages 424-426.

[18] João Carreira et Andrew Zisserman. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". In : *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pages 4724-4733.

[19] Turgay Çelik et Kai-Kuang Ma. "Multitemporal Image Change Detection Using Undecimated Discrete Wavelet Transform and Active Contours". *IEEE Transactions on Geoscience and Remote Sensing* 49.2 (2011), pages 706-716.

[20] Siddhartha Chandra, Camille Couprie et Iasonas Kokkinos. “Deep Spatio-Temporal Random Fields for Efficient Video Segmentation”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2018, pages 8915-8924.

[21] Aditya Chattopadhyay et al. “Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks”. In : *IEEE Winter Conference on Applications of Computer Vision*. IEEE Computer Society, 2018, pages 839-847.

[22] Ming Cheng, Kunjing Cai et Ming Li. “RWF-2000: An Open Large Scale Video Database for Violence Detection”. In : *International Conference on Pattern Recognition*. IEEE, 2020, pages 4183-4190.

[23] Maximilian Christ et al. “Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)”. *Neurocomputing* 307 (2018), pages 72-77. issn : 0925-2312.

[24] Dorin Comaniciu et Peter Meer. “Mean Shift: A Robust Approach Toward Feature Space Analysis”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5 (2002), pages 603-619.

[25] P. Coppin et al. “Digital change detection methods in ecosystem monitoring: A review”. *Remote Sensing* 25.5 (2004), pages 1565-1596.

[26] Yady Tatiana Solano Correa et al. “A Method for the Analysis of Small Crop Fields in Sentinel-2 Dense Time Series”. *IEEE transactions on geoscience and remote sensing* 58.3 (2020), pages 2150-2164.

[27] Xavier Cortés, Donatello Conte et Hubert Cardot. “A new bag of visual words encoding method for human action recognition”. In : *International Conference on Pattern Recognition*. IEEE Computer Society, 2018, pages 2480-2485.

[28] Dubravko Culibrk et Nicu Sebe. “Temporal Dropout of Changes Approach to Convolutional Learning of Spatio-Temporal Features”. In : *ACM International Conference on Multimedia*. Sous la direction de Kien A. Hua et al. ACM, 2014, pages 1201-1204.

[29] Xinyan Dai et al. “Norm-Explicit Quantization: Improving Vector Quantization for Maximum Inner Product Search”. In : *Innovative Applications of Artificial Intelligence*. AAAI Press, 2020, pages 51-58.

[30] Navneet Dalal, Bill Triggs et Cordelia Schmid. “Human Detection Using Oriented Histograms of Flow and Appearance”. In : *European Conference on Computer Vision, Part II*. Sous la direction d’Ales Leonardis, Horst Bischof et Axel Pinz. Tome 3952. Lecture Notes in Computer Science. Springer, 2006, pages 428-441.

[31] Jia Deng et al. “ImageNet: A large-scale hierarchical image database”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2009, pages 248-255.

[32] Dawa Derksen, Jordi Inglada et Julien Michel. "Geometry Aware Evaluation of Handcrafted Superpixel-Based Features and Convolutional Neural Networks for Land Cover Mapping Using Satellite Imagery". *Remote Sensing* 12.3 (2020), page 513.

[33] Matthias Drusch et al. "Sentinel-2: ESA's optical high-resolution mission for GMES operational services". *Remote sensing of Environment* 120 (2012), pages 25-36.

[34] Nicola Falco et al. "Change Detection in VHR Images Based on Morphological Attribute Profiles". *IEEE Geoscience and Remote Sensing Letters* 10.3 (2013), pages 636-640.

[35] Haoshu Fang et al. "RMPE: Regional Multi-person Pose Estimation". In : *IEEE International Conference on Computer Vision*. IEEE Computer Society, 2017, pages 2353-2362.

[36] Gunnar Farnebäck. "Two-Frame Motion Estimation Based on Polynomial Expansion". In : *Scandinavian Conference on Image Analysis*. Sous la direction de Josef Bigün et Tomas Gustavsson. Tome 2749. Lecture Notes in Computer Science. Springer, 2003, pages 363-370.

[37] Hassan Ismail Fawaz et al. "Deep learning for time series classification: a review". *Data Mining and Knowledge Discovery* 33.4 (2019), pages 917-963.

[38] Christoph Feichtenhofer, Axel Pinz et Richard P. Wildes. "Spatiotemporal Multiplier Networks for Video Action Recognition". In : *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2017, pages 7445-7454.

[39] Christoph Feichtenhofer, Axel Pinz et Andrew Zisserman. "Convolutional Two-Stream Network Fusion for Video Action Recognition". In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2016, pages 1933-1941.

[40] Hajer Fradi et Jean-Luc Dugelay. "Spatial and temporal variations of feature tracks for crowd behavior analysis". *Journal on Multimodal User Interfaces* 10.4 (2016), pages 307-317.

[41] Herbert Freeman et Larry S. Davis. "A Corner-Finding Algorithm for Chain-Coded Curves". *IEEE Transactions on Computers* 26.3 (1977), pages 297-303.

[42] Vivien Sainte Fare Garnot et al. "Satellite Image Time Series Classification With Pixel-Set Encoders and Temporal Self-Attention". In : *IEEE Computer Vision and Pattern Recognition*. IEEE, 2020, pages 12322-12331.

[43] Solomon W. Golomb. "Run-length encodings (Corresp.)" *IEEE Trans. Inf. Theory* 12.3 (1966), pages 399-401.

[44] Ross Goroshin et al. "Unsupervised Feature Learning from Temporal Data". In : *International Conference on Learning Representations*. Sous la direction d'Yoshua Bengio et Yann LeCun. 2015.

[45] Leo Grady. “Multilabel Random Walker Image Segmentation Using Prior Models”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2005, pages 763-770.

[46] Alex Graves et Jürgen Schmidhuber. “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”. *Neural Networks* 18.5-6 (2005), pages 602-610.

[47] Yi-Lan Guo, Ji-Xiang Du et Chuan-Min Zhai. “Event Recognition Based on a Local Space-Time Interest Points and Self-Organization Feature Map Method”. In : *Advanced Intelligent Computing*. Sous la direction de De-Shuang Huang et al. Tome 6838. Lecture Notes in Computer Science. Springer, 2011, pages 242-249.

[48] Olivier Hagolle et al. “A Multi-Temporal and Multi-Spectral Method to Estimate Aerosol Optical Thickness over Land, for the Atmospheric Correction of FormoSat-2, LandSat, VENμS and Sentinel-2 Images”. *Remote Sensing* 7.3 (2015), pages 2668-2691.

[49] Tal Hassner, Yossi Itcher et Orit Kliper-Gross. “Violent flows: Real-time detection of violent crowd behavior”. In : *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE Computer Society, 2012, pages 1-6.

[50] Nima Hatami, Yann Gavet et Johan Debayle. “Classification of time-series images using deep convolutional neural networks”. In : *International Conference on Machine Vision*. Sous la direction d’Antanas Verikas et al. Tome 10696. SPIE Proceedings. SPIE, 2017, 106960Y.

[51] Kaiming He et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In : *IEEE International Conference on Computer Vision*. IEEE Computer Society, 2015, pages 1026-1034.

[52] Sepp Hochreiter et al. “Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies”. In : *IEEE A Field Guide to Dynamical Recurrent Networks*. 2001.

[53] Nesma Houmani et al. “On-line Signature Verification on a Mobile Platform”. In : *Mobile Computing, Applications, and Services*. Sous la direction de Martin L. Griss et Guang Yang. Tome 76. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer, 2010, pages 396-400.

[54] Gang Hu et Qigang Gao. “Dynamic Perceptual Attribute-Based Hidden Conditional Random Fields for Gesture Recognition”. In : *International Conference on Image Analysis and Recognition*. Sous la direction de Mohamed Kamel et Aurélio J. C. Campilho. Tome 9164. Lecture Notes in Computer Science. Springer, 2015, pages 259-268.

[55] Bohao Huang et al. “Large-Scale Semantic Classification: Outcome of the First Year of Inria Aerial Image Labeling Benchmark”. In : *IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pages 6947-6950.

[56] Shih-Shinh Huang et al. “Combining Histograms of Oriented Gradients with Global Feature for Human Detection”. In : *Advances in Multimedia Modeling, Part II*. Tome 6524. Lecture Notes in Computer Science. Springer, 2011, pages 208-218.

[57] David A. Huffman. “A Method for the Construction of Minimum-Redundancy Codes”. *Proceedings of the IRE* 40.9 (1952), pages 1098-1101.

[58] F.N. Iandola et al. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size”. *Computing Research Repository* abs/1602.07360 (2016).

[59] Dino Ienco et al. “Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks”. *IEEE Geoscience and Remote Sensing Letters* 14.10 (2017), pages 1685-1689.

[60] Roberto Interdonato et al. “DuPLO: A DUal view Point deep Learning architecture for time series classificatiOn”. *CoRR* abs/1809.07589 (2018).

[61] Sergey Ioffe et Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In : *International Conference on Machine Learning*. Sous la direction de Francis R. Bach et David M. Blei. Tome 37. JMLR Workshop and Conference Proceedings. JMLR.org, 2015, pages 448-456.

[62] John R. Jensen. “Urban change detection mapping using Landsat digital data”. *Cartography and Geographic Information Science* 8.21 (1981), pages 127-147.

[63] Shuiwang Ji et al. “3D Convolutional Neural Networks for Human Action Recognition”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pages 221-231.

[64] Yuan Jing, Jinshan Hao et Peng Li. “Learning Spatiotemporal Features of CSI for Indoor Localization With Dual-Stream 3D Convolutional Neural Networks”. *IEEE Access* 7 (2019), pages 147571-147585.

[65] R.D. Johnson et E.S. Kasischke. “Change vector analysis: A technique for the multispectral monitoring of land cover and condition”. *Remote Sensing* 19.16 (1998), pages 411-426.

[66] Ekaterina Kalinicheva, Jérémie Sublime et Maria Trocan. “Unsupervised Satellite Image Time Series Clustering Using Object-Based Approaches and 3D Convolutional Autoencoder”. *Remote Sensing* 12.11 (2020), page 1816.

[67] Andrej Karpathy et al. “Large-Scale Video Classification with Convolutional Neural Networks”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2014, pages 1725-1732.

[68] David J. Ketchen et Christopher L. Shool. “The application of cluster analysis in strategic managment research: an analysis and critique”. *Strategic Management Journal* 17.6 (1996), pages 441-458.

[69] Lynda Khiali et al. “Detection of spatio-temporal evolutions on multi-annual satellite image time series: A clustering based approach”. *International Journal of Applied Earth Observation and Geoinformation* 74 (2019), pages 103-119.

[70] Ender Konukoglu et al. “Towards an Identification of Tumor Growth Parameters from Time Series of Images”. In : *Medical Image Computing and Computer-Assisted Intervention, Part I*. Sous la direction de Nicholas Ayache, Sébastien Ourselin et Anthony J. Maeder. Tome 4791. Lecture Notes in Computer Science. Springer, 2007, pages 549-556.

[71] Okan Köpüklü et al. “Resource Efficient 3D Convolutional Neural Networks”. In : *IEEE International Conference on Computer Vision Workshops*. IEEE, 2019, pages 1910-1919.

[72] Adriana Kovashka et Kristen Grauman. “Learning a hierarchy of discriminative space-time neighborhood features for human action recognition”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2010, pages 2046-2053.

[73] Alex Krizhevsky, Ilya Sutskever et Geoffrey E. Hinton. “ImageNet classification with deep convolutional neural networks”. *Communications ACM* 60.6 (2017), pages 84-90.

[74] Eric F. Lambin et Alan H. Strahlers. “Change-vector analysis in multitemporal space: A tool to detect and categorize land-cover change processes using high temporal-resolution satellite data”. *Remote Sensing of Environment* 48.2 (1994), pages 231-244.

[75] Michele Linardi et al. “VALMOD: A Suite for Easy and Exact Detection of Variable Length Motifs in Data Series”. In : *Special Interest Group On Management of Data, SIGMOD*. Sous la direction de Gautam Das, Christopher M. Jermaine et Philip A. Bernstein. ACM, 2018, pages 1757-1760.

[76] Chunhui Liu et al. “PKU-MMD: A Large Scale Benchmark for Skeleton-Based Human Action Understanding”. In : *Workshop on Visual Analysis in Smart and Connected Communities*. Sous la direction de Xiaobai Liu et al. ACM, 2017, pages 1-8.

[77] Mengyuan Liu et al. “Salient pairwise spatio-temporal interest points for real-time activity recognition”. *CAAI Transactions on Intelligence Technology* 1.1 (2016), pages 14-29.

[78] Stuart P. Lloyd. “Least squares quantization in PCM”. *IEEE Transactions on Information Theory* 28.2 (1982), pages 129-136.

[79] Xiangjun Lu et al. “Pay Attention to Raw Traces: A Deep Learning Architecture for End-to-End Profiling Attacks”. *IACR transactions on cryptographic hardware and embedded systems* 2021.3 (2021), pages 235-274.

[80] Van-Bao Ly, Sonia Garcia-Salicetti et Bernadette Dorizzi. “On Using the Viterbi Path Along With HMM Likelihood Information for Online Signature Verification”. *IEEE Transactions Systems, Man and Cybernetics, Part B* 37.5 (2007), pages 1237-1247.

[81] Chih-Yao Ma et al. “TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition”. *Signal Processing: Image Communication* 71 (2019), pages 76-87.

[82] Tara Madhyastha et al. “Current methods and limitations for longitudinal fMRI analysis across development”. *Develop Cogn Neuro* 33 (2018), pages 118-128.

[83] Norbert Marwan et al. “Recurrence plots for the analysis of complex systems”. *Physics Reports* 438.5 (2007), pages 237-329.

[84] Nicola Di Mauro et al. “End-to-end Learning of Deep Spatio-temporal Representations for Satellite Image Time Series Classification”. In : *Discovery Challenges co-located with European Conference on Machine Learning - Principle and Practice of Knowledge Discovery in Database*. Sous la direction de Roberto Corizzo et Dino Ienco. Tome 1972. CEUR Workshop Proceedings. CEUR-WS.org, 2017.

[85] Nicolas Méger et al. “Ranking evolution maps for Satellite Image Time Series exploration: application to crustal deformation and environmental monitoring”. *Data Mining and Knowledge Discovery* 33.1 (2019), pages 131-167.

[86] Magdi A. Mohamed et Paul D. Gader. “Handwritten Word Recognition Using Segmentation-Free Hidden Markov Modeling and Segmentation-Based Dynamic Programming Techniques”. *IEEE Transactions Pattern Analysis and Machine Intelligence* 18.5 (1996), pages 548-554.

[87] Gruffydd Morris et Plamen Angelov. “Real-time novelty detection in video using background subtraction techniques: State of the art a practical review”. In : *Systems, Man, and Cybernetics, SMC*. IEEE, 2014, pages 537-543.

[88] Roopal Nahar, Akanksha Baranwal et K. Madhava Krishna. “FPGA based parallelized architecture of efficient graph based image segmentation algorithm”. In : *IEEE International Conference on Robotics and Biomimetics*. IEEE, 2017, pages 98-103.

[89] Farhood Negin et al. “A Decision Forest Based Feature Selection Framework for Action Recognition from RGB-Depth Cameras”. In : *International Conference on Image Analysis and Recognition*. Sous la direction de Mohamed Kamel et Aurélio J. C. Campilho. Tome 7950. Lecture Notes in Computer Science. Springer, 2013, pages 648-657.

[90] Giap Nguyen, Patrick Franco et Jean-Marc Ogier. “Space-Filling Curve for Image Dynamical Indexing”. In : *CInternational Symposium on Computer and Information Sciences*. Sous la direction d’Erol Gelenbe et Ricardo Lent. Springer, 2012, pages 311-319.

[91] Enrique Bermejo Nievas et al. “Violence Detection in Video Using Computer Vision Techniques”. In : *Computer Analysis of Images and Patterns, Part II*. Sous la direction de Pedro Real et al. Tome 6855. Lecture Notes in Computer Science. Springer, 2011, pages 332-339.

[92] Alexandre Ninassi et al. “Considering Temporal Variations of Spatial Visual Distortions in Video Quality Assessment”. *IEEE Journal of Selected Topics in Signal Processing* 3.2 (2009), pages 253-265.

[93] Shibli Nisar, Omar Usman Khan et Muhammad Tariq. “An Efficient Adaptive Window Size Selection Method for Improving Spectrogram Visualization”. *Computational Intelligence and Neuroscience* 2016 (2016), 6172453:1-6172453:13.

[94] Abraham Montoya Obeso et al. “Forward-backward visual saliency propagation in Deep NNs vs internal attentional mechanisms”. In : *International Conference on Image Processing Theory, Tools and Applications*. IEEE, 2019, pages 1-6.

[95] Eshed Ohn-Bar et Mohan M. Trivedi. “Joint Angles Similarities and HOG2 for Action Recognition”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2013, pages 465-470.

[96] Mayu Otani et al. “Rethinking the Evaluation of Video Summaries”. In : *IEEE Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 2019, pages 7596-7604.

[97] Yagya Raj Pandeya et Joonwhoan Lee. “Deep learning-based late fusion of multimodal information for emotion classification of music video”. *Multimedia Tools and Applications* 80.2 (2021), pages 2887-2905.

[98] Charlotte Pelletier, Geoffrey I. Webb et François Petitjean. “Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series”. *Remote Sensing* 11.5 (2019), page 523.

[99] François Petitjean et Jonathan Weber. “Efficient Satellite Image Time Series Analysis Under Time Warping”. *IEEE Geoscience and Remote Sensing Letters* 11.6 (2014), pages 1143-1147.

[100] François Petitjean et al. “Spatio-temporal reasoning for the classification of satellite image time series”. *Pattern Recognition Letters* 33.13 (2012), pages 1805-1815.

[101] François Petitjean. “Dynamic Time Warping : Apports théoriques pour l’analyse de données temporelles. Application à la classification de séries temporelles d’images satellites”. Thèse de doctorat. Université de Strasbourg, sept. 2012.

[102] Charles Ruizhongtai Qi et al. “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”. In : *Advances in Neural Information Processing Systems*. Sous la direction d’Isabelle Guyon et al. 2017, pages 5099-5108.

[103] Charles Ruizhongtai Qi et al. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2017, pages 77-85.

[104] Zhaofan Qiu, Ting Yao et Tao Mei. “Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks”. In : *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pages 5534-5542.

[105] L.R. Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. *Proceedings of the IEEE* 77.2 (1989), pages 257-286.

[106] Chotirat (Ann) Ratanamahatana et al. “A Novel Bit Level Time Series Representation with Implication of Similarity Search and Clustering”. In : *Advances in Knowledge Discovery and Data Mining*. Sous la direction de Tu Bao Ho, David Wai-Lok Cheung et Huan Liu. Tome 3518. Lecture Notes in Computer Science. Springer, 2005, pages 771-777.

[107] Wei Ren et al. “State-of-the-art on spatio-temporal information-based video retrieval”. *Pattern Recognition* 42.2 (2009), pages 267-282.

[108] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. *Journal of Computational and Applied Mathematics* 20 (1987), pages 53-65. issn : 0377-0427.

[109] Alejandro Pasos Ruiz et al. “The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances”. *Data Mining and Knowledge Discovery* 35.2 (2021), pages 401-449.

[110] Olga Russakovsky et al. “ImageNet Large Scale Visual Recognition Challenge”. *International Journal of Computer Vision* 115.3 (2015), pages 211-252.

[111] Marc Russwurm et Marco Körner. “Convolutional LSTMs for Cloud-Robust Segmentation of Remote Sensing Imagery”. *CoRR* abs/1811.02471 (2018).

[112] Marc Russwurm et Marco Körner. “Temporal Vegetation Modelling Using Long Short-Term Memory Networks for Crop Identification from Medium-Resolution Multi-spectral Satellite Images”. In : *IEEE Computer Vision and Pattern Recognition Workshops*. IEEE Computer Society, 2017, pages 1496-1504.

[113] Radu Bogdan Rusu et Steve Cousins. “3D is here: Point Cloud Library (PCL)”. In : *IEEE International Conference on Robotics and Automation*. IEEE, 2011.

[114] Samy Sadek et al. “Real-Time Automatic Traffic Accident Recognition Using HFG”. In : *International Conference on Pattern Recognition*. IEEE Computer Society, 2010, pages 3348-3351.

[115] Paul Scovanner, Saad Ali et Mubarak Shah. “A 3-dimensional sift descriptor and its application to action recognition”. In : *International Conference on Multimedia*. Sous la direction de Rainer Lienhart et al. ACM, 2007, pages 357-360.

[116] Mouna Selmi et Mounim A. El-Yacoubi. “Multimodal Sequential Modeling and Recognition of Human Activities”. In : *International Conference Computers Helping People with Special Needs*. Sous la direction de Klaus Miesenberger, Christian Bühler et Petr Penáz. Tome 9759. Lecture Notes in Computer Science. Springer, 2016, pages 541-548.

[117] Mouna Selmi, Mounim A. El-Yacoubi et Bernadette Dorizzi. “On the sensitivity of spatio-temporal interest points to person identity”. In : *IEEE Southwest Symposium on Image Analysis and Interpretation*. IEEE Computer Society, 2012, pages 69-72.

[118] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. *International Journal of Computer Vision* 128.2 (2020), pages 336-359.

[119] C. Senf et al. “Mapping land cover in complex Mediterranean landscapes using Landsat: Improved classification accuracies from integrating multi-seasonal and synthetic imagery”. *Remote Sensing of Environment* 156 (2015), pages 527-536.

[120] Atsushi Shimada, Hajime Nagahara et Rin-Ichiro Taniguchi. “Change detection on light field for active video surveillance”. In : *IEEE Advanced Video and Signal Based Surveillance*. IEEE Computer Society, 2015, pages 1-6.

[121] Connor Shorten et Taghi M. Khoshgoftaar. “A survey on Image Data Augmentation for Deep Learning”. *Journal on Big Data* 6 (2019), page 60.

[122] Jamie Shotton et al. “Real-Time Human Pose Recognition in Parts from Single Depth Images”. In : *Machine Learning for Computer Vision*. Sous la direction de Roberto Cipolla, Sebastiano Battiato et Giovanni Maria Farinella. Tome 411. Studies in Computational Intelligence. Springer, 2013, pages 119-135.

[123] Cristian Sminchisescu, Atul Kanaujia et Dimitris N. Metaxas. “Conditional models for contextual human motion recognition”. *Computer Vision and Image Understanding* 104.2-3 (2006), pages 210-220.

[124] Pierre Soille. “Constrained Connectivity for Hierarchical Image Decomposition and Simplification”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.7 (2008), pages 1132-1145.

[125] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. *Journal of Machine Learning Research* 15.1 (2014), pages 1929-1958.

[126] Andrei Stoian et al. “Land Cover Maps Production with High Resolution Satellite Image Time Series and Convolutional Neural Networks: Adaptations and Limits for Operational Systems”. *Remote Sensing* 11.17 (2019), page 1986.

[127] Nico Stuurman et Ronald D. Vale. “Impact of New Camera Technologies on Discoveries in Cell Biology”. *Biol Bull* 231.1 (2016), pages 5-13.

[128] Yukun Su et al. “Human Interaction Learning on 3D Skeleton Point Clouds for Video Violence Recognition”. In : *European Conference on Computer Vision*. Sous la direction d’Andrea Vedaldi et al. Tome 12349. Lecture Notes in Computer Science. Springer, 2020, pages 74-90.

[129] Neil Sumpter et Andrew J. Bulpitt. “Learning spatio-temporal patterns for predicting object behaviour”. *Image and Vision Computing* 18.9 (2000), pages 697-704.

[130] Guoliang Tang, ZhiJing Liu et Jing Xiong. “Distinctive image features from illumination and scale invariant keypoints”. *Multimedia Tools and Applications* 78.16 (2019), pages 23415-23442.

[131] Ludvík Tesar et al. “Medical image analysis of 3D CT images based on extension of Haralick texture features”. *Computerized Medical Imaging and Graphics* 32.6 (2008), pages 513-520.

[132] Hans G. Tillmann et Jessica Siddins. “The "bonn connection" and its consequences: Paul Menzerath and Werner Meyer-Eppler’s reunification of phonetics and phonology and the emergence of a new phonetic speech science based on Shannon’s mathematical theory of communication”. In : *International Workshop on the History of Speech Communication Research*. Sous la direction de Rüdiger Hoffmann et Jürgen Trouvain. ISCA, 2015, pages 128-139.

[133] Du Tran et al. “A Closer Look at Spatiotemporal Convolutions for Action Recognition”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2018, pages 6450-6459.

[134] Çaglayan Tuna. “Morphological Hierarchies for Satellite Image Time Series”. Theses. Université de Bretagne Sud, déc. 2020.

[135] Raviteja Vemulapalli et Rama Chellappa. “Rolling Rotations for Recognizing Human Actions from 3D Skeletal Data”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2016, pages 4471-4479.

[136] J. Verbesselt et al. “Detecting trend and seasonal changes in satellite image time series”. *Remote Sensing of Environment* 114.1 (2010), pages 106-115.

[137] Haofan Wang et al. “Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE, 2020, pages 111-119.

[138] Jiang Wang et al. “Mining actionlet ensemble for action recognition with depth cameras”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2012, pages 1290-1297.

[139] Limin Wang et al. “Temporal Segment Networks: Towards Good Practices for Deep Action Recognition”. In : *European Conference on Computer Vision*. Sous la direction de Bastian Leibe et al. Tome 9912. Lecture Notes in Computer Science. Springer, 2016, pages 20-36.

[140] Pichao Wang et al. “Scene Flow to Action Map: A New Representation for RGB-D Based Action Recognition with Convolutional Neural Networks”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2017, pages 416-425.

[141] Yue Wang et al. “Dynamic Graph CNN for Learning on Point Clouds”. *ACM Transactions on Graphics* 38.5 (2019), 146:1-146:12.

[142] Zhiguang Wang et Tim Oates. “Imaging Time-Series to Improve Classification and Imputation”. In : *International Joint Conference on Artificial Intelligence*. Sous la direction de Qiang Yang et Michael J. Wooldridge. AAAI Press, 2015, pages 3939-3945.

[143] Jonathan Weber, Sébastien Lefèvre et Pierre Gançarski. “Spatio-temporal Quasi-Flat Zones for Morphological Video Segmentation”. In : *Mathematical Morphology and Its Applications to Image and Signal Processing*. Sous la direction de Pierre Soille, Martino Pesaresi et Georgios K. Ouzounis. Tome 6671. Lecture Notes in Computer Science. Springer, 2011, pages 178-189.

[144] Terry A. Welch. “A Technique for High-Performance Data Compression”. *Computer* 17.6 (1984), pages 8-19.

[145] Junwu Weng et al. “Discriminative Spatio-Temporal Pattern Discovery for 3D Action Recognition”. *IEEE Transactions on Circuits and Systems for Video Technology*. 29.4 (2019), pages 1077-1089.

[146] Tracy L. Westeyn et al. “Georgia tech gesture toolkit: supporting experiments in gesture recognition”. In : *PInternational Conference on Multimodal Interfaces*. Sous la direction de Sharon L. Oviatt et al. ACM, 2003, pages 85-92.

[147] Martin Wöllmer et al. “Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling”. In : *INTERSPEECH Annual Conference of the International Speech Communication Association*. Sous la direction de Takao Kobayashi, Keikichi Hirose et Satoshi Nakamura. ISCA, 2010, pages 2362-2365.

[148] Wenxuan Wu, Zhongang Qi et Fuxin Li. “PointConv: Deep Convolutional Networks on 3D Point Clouds”. In : *IEEE Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 2019, pages 9621-9630.

[149] Long Xu et al. “Violent video detection based on MoSIFT feature and sparse coding”. In : *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pages 3538-3542.

[150] S. S. Young et C. Y. Wang. “Land-cover change analysis of China using global-scale Pathfinder AVHRR Landcover (PAL) data, 1982?92”. *Remote Sensing* 22.8 (2001), pages 1457-1477.

[151] Bolei Zhou et al. “Learning Deep Features for Discriminative Localization”. In : *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2016, pages 2921-2929.

[152] Xi Zhou et al. “SIFT-Bag kernel for video event analysis”. In : *International Conference on Multimedia*. Sous la direction d’Abdulmotaleb El-Saddik et al. ACM, 2008, pages 229-238.

[153] Mohammadreza Zolfaghari, Kamaljeet Singh et Thomas Brox. “ECO: Efficient Convolutional Network for Online Video Understanding”. In : *European Conference on Computer Vision*. Sous la direction de Vittorio Ferrari et al. Tome 11206. Lecture Notes in Computer Science. Springer, 2018, pages 713-730.