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

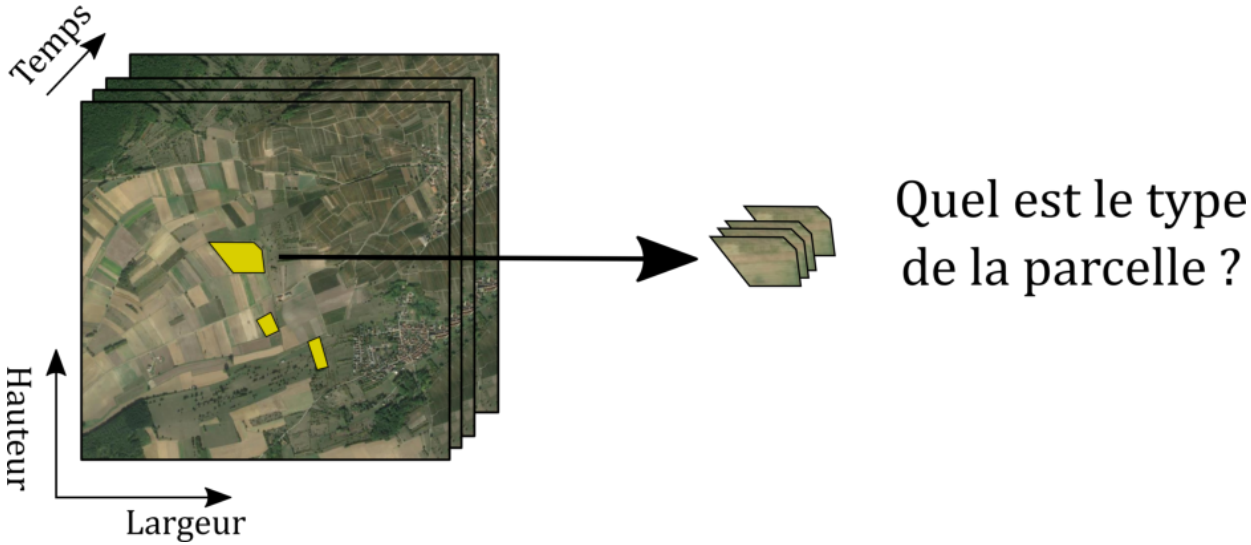
Mohamed CHELALI

26 novembre 2021

Université de Paris

Projet ANR : TIMES

- Objectif : analyse des changements environnementaux
- Motivation :
  - Exploitation de masses de données (visuelles) hétérogènes à haute fréquence temporelle



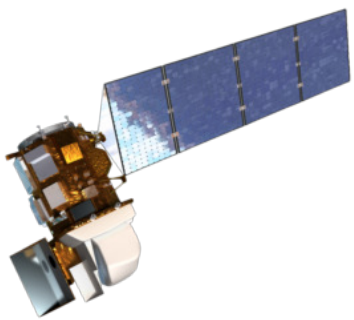
Plan

Introduction

Université de Paris

Contexte

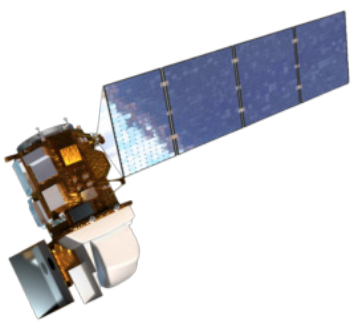
- Évolution de la technologie

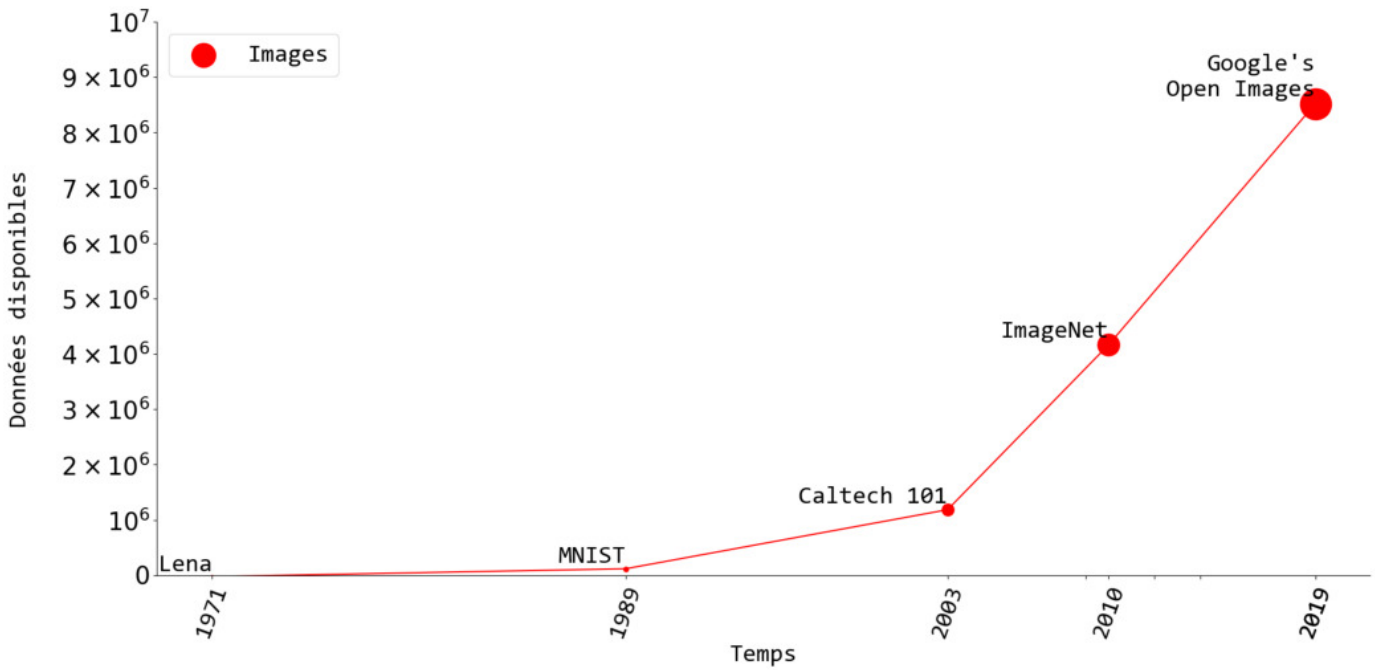


Université de Paris

Contexte

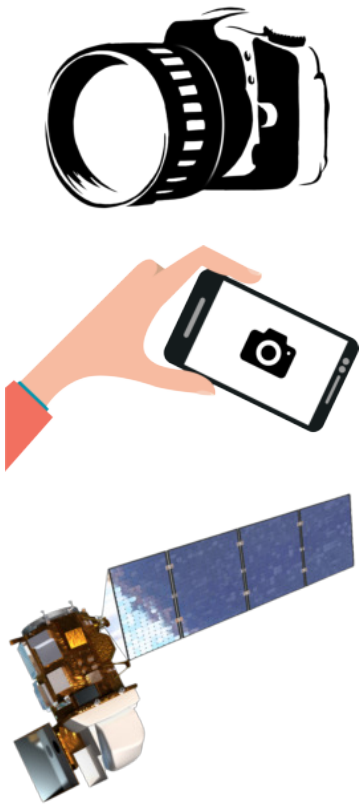
- Évolution de la technologie
- Augmentation de **la quantité d'images et de vidéos ou de séquences temporelles d'images (STI)**

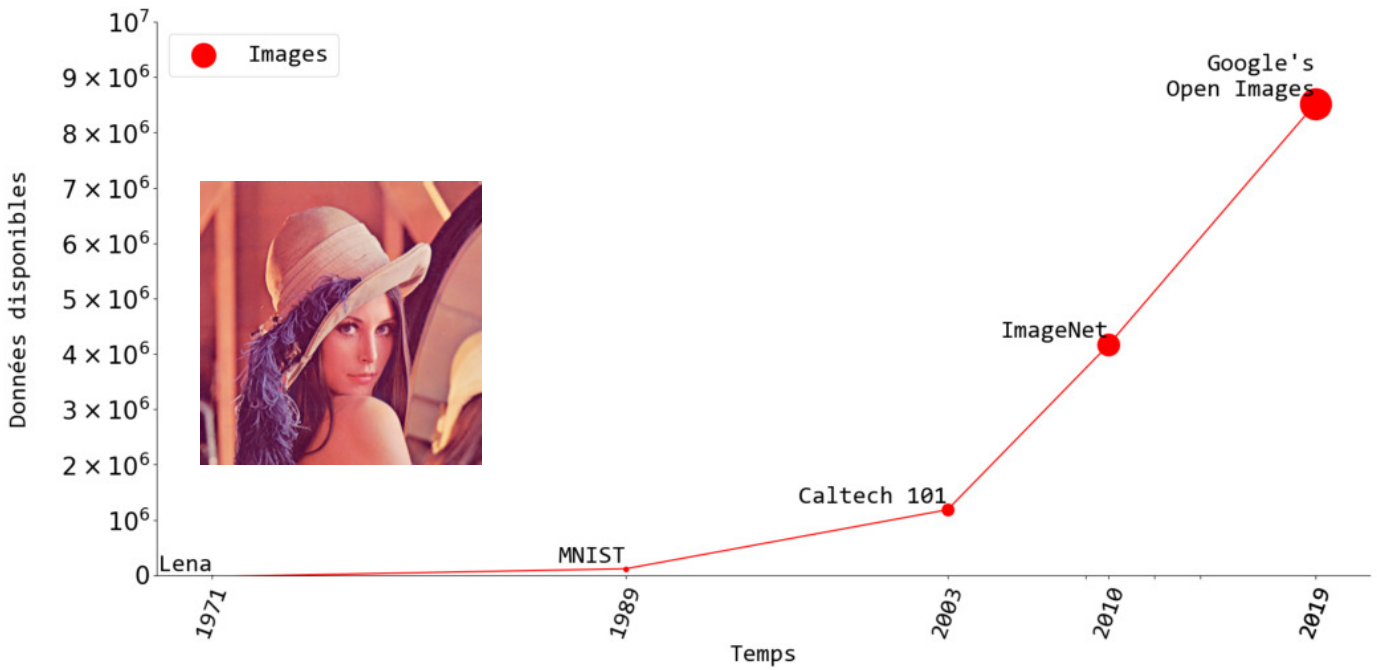




Contexte

- Évolution de la technologie
- Augmentation de **la quantité d'images et de vidéos ou de séquences temporelles d'images (STI)**

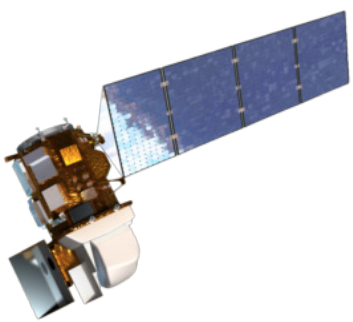


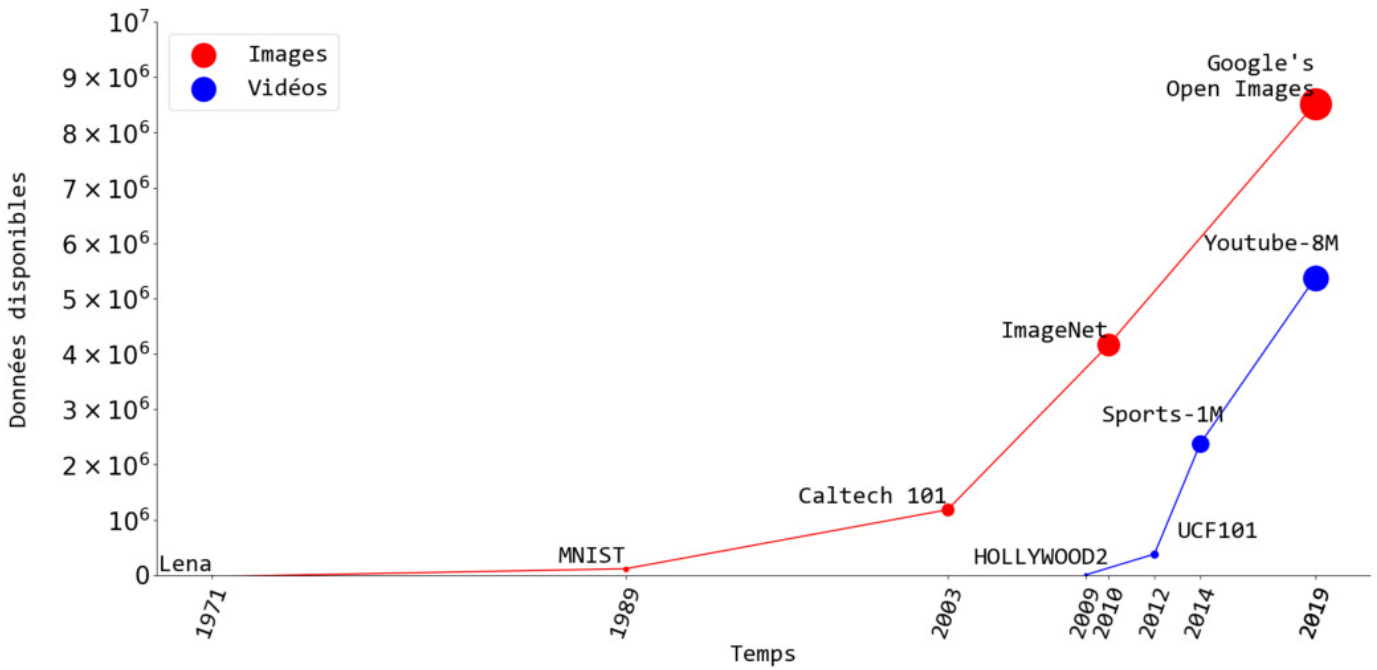


Université de Paris

Contexte

- Évolution de la technologie
- Augmentation de **la quantité d'images et de vidéos ou de séquences temporelles d'images (STI)**





Université de Paris

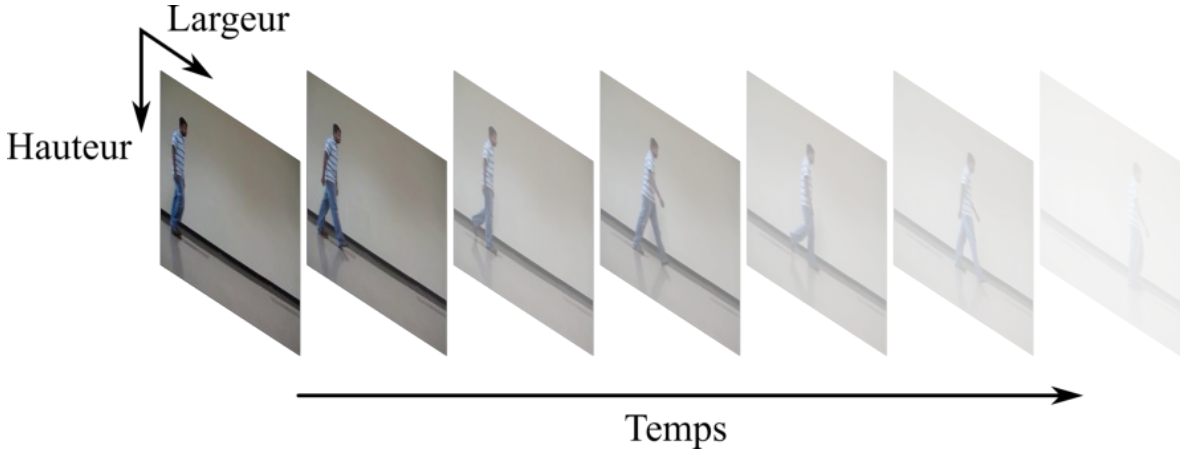
Contexte

- Données initiales des séquences temporelles d'images (STI) : $2D + t$

Université de Paris

Contexte

- Données initiales des séquences temporelles d'images (STI) : $2D + t$



- Étudier l'information du domaine spatial au cours du temps

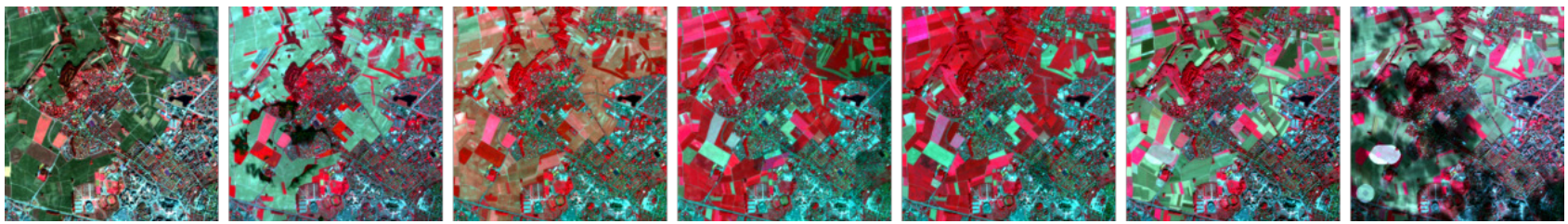
Université de Paris

Contexte

- Données initiales des séquences temporelles d'images (STI) : $2D + t$
- Étudier l'information du domaine spatial au cours du temps
- Deux types de STI :
  - **Séquence d'images avec une continuité visuelle**
  -

Contexte

- Données initiales des séquences temporelles d'images (STI) : $2D + t$
- Étudier l'information du domaine spatial au cours du temps
- Deux types de STI :
  - **Séquence d'images avec une continuité visuelle**
  - **Séquence avec des images ponctuelles**



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

Université de Paris

Problématiques

Séquence d'images avec une continuité visuelle

- La caméra peut être fixe ou en mouvement
- Vitesse de déplacement (caméra ou objet dans la scène)

(a) Caméra fixe

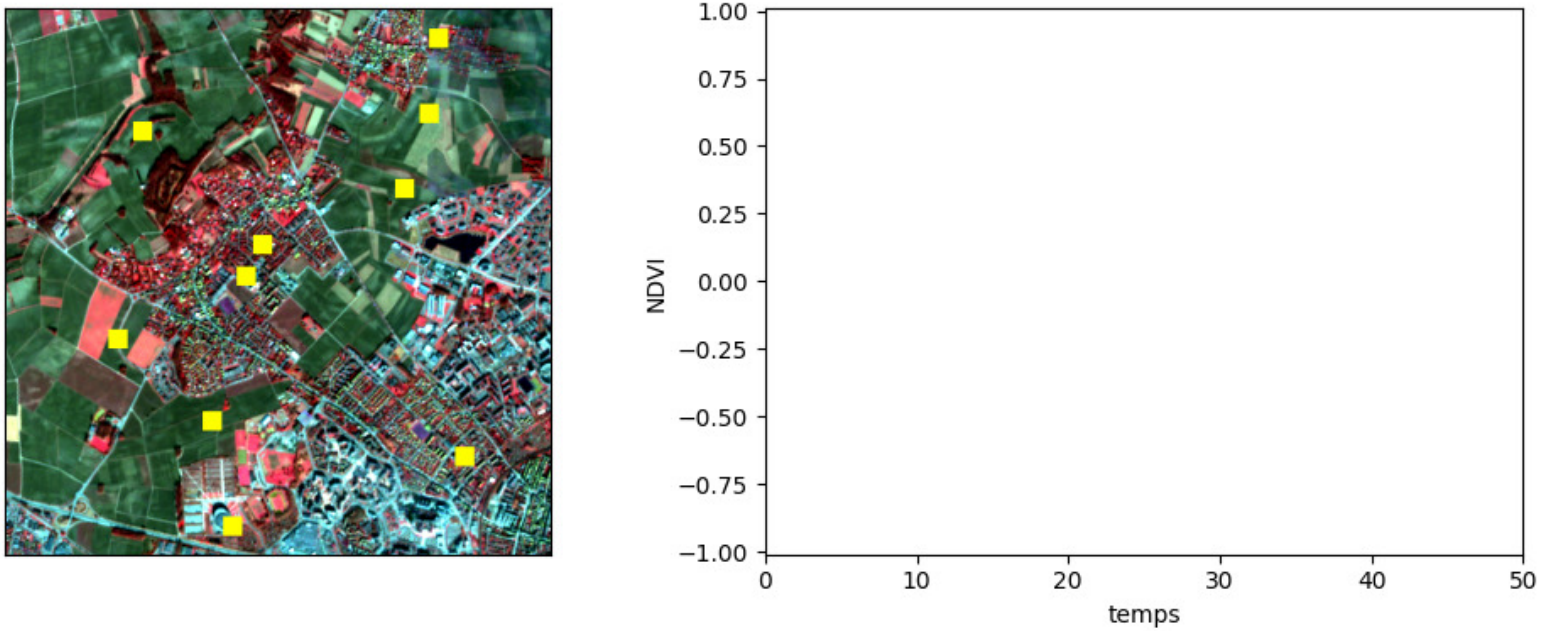
(b) Caméra en mouvement

Université de Paris

Problématiques

Séquence d'images avec un contenu non-déformable

- La caméra est toujours fixe sur la même scène
- Étude de l'évolution temporelle

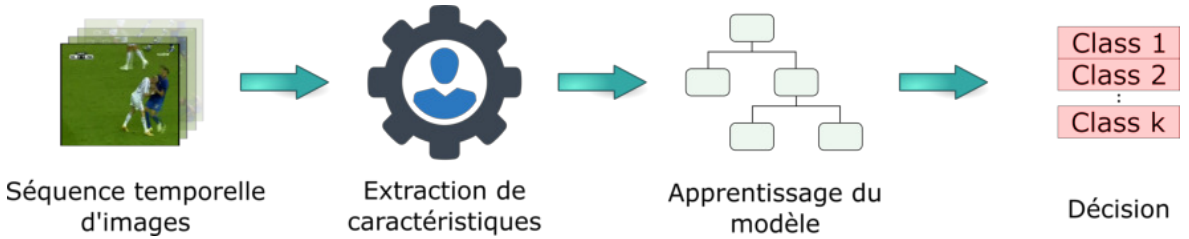


(a) Caméra fixe

Méthodes de l'état de l'art pour l'analyse des séquences temporelles d'images

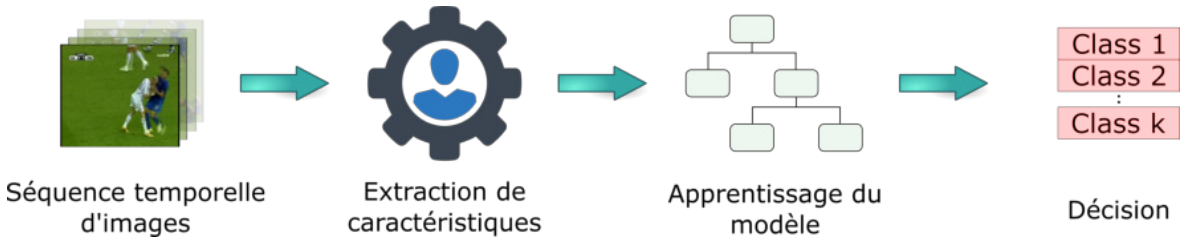
Types des caractéristiques

- Caractéristiques expertes (artisanales ou *hand-crafted*)

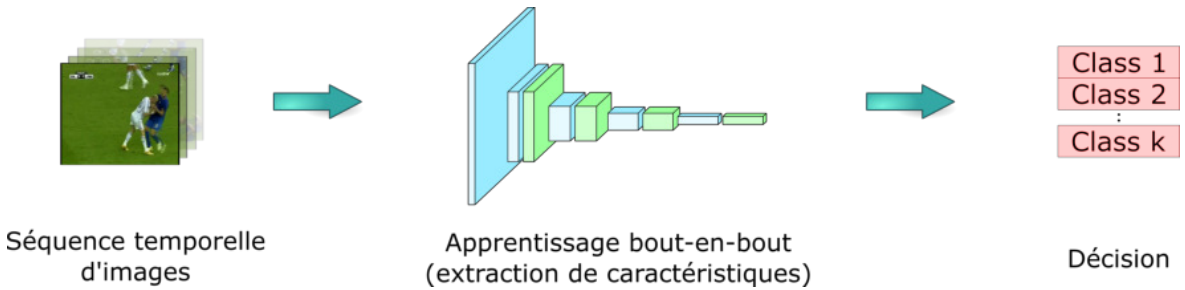


Types des caractéristiques

- Caractéristiques expertes (artisanales ou *hand-crafted*)

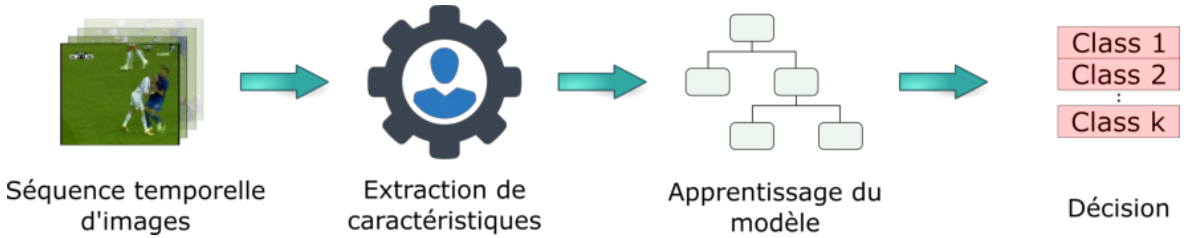


- Architecture expertes avec des caractéristiques apprises (CNN, RNN)

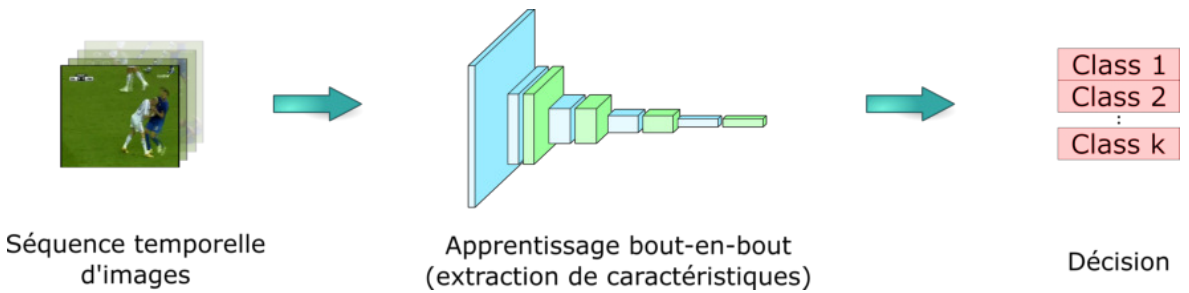


Types des caractéristiques

- Caractéristiques expertes (artisanales ou *hand-crafted*)



- Architecture expertes avec des caractéristiques apprises (CNN, RNN)

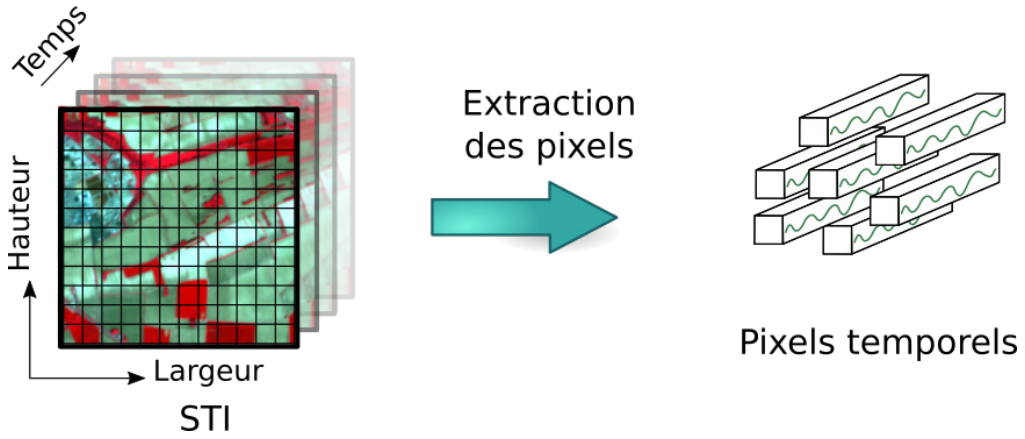


Artisanales
généralistes

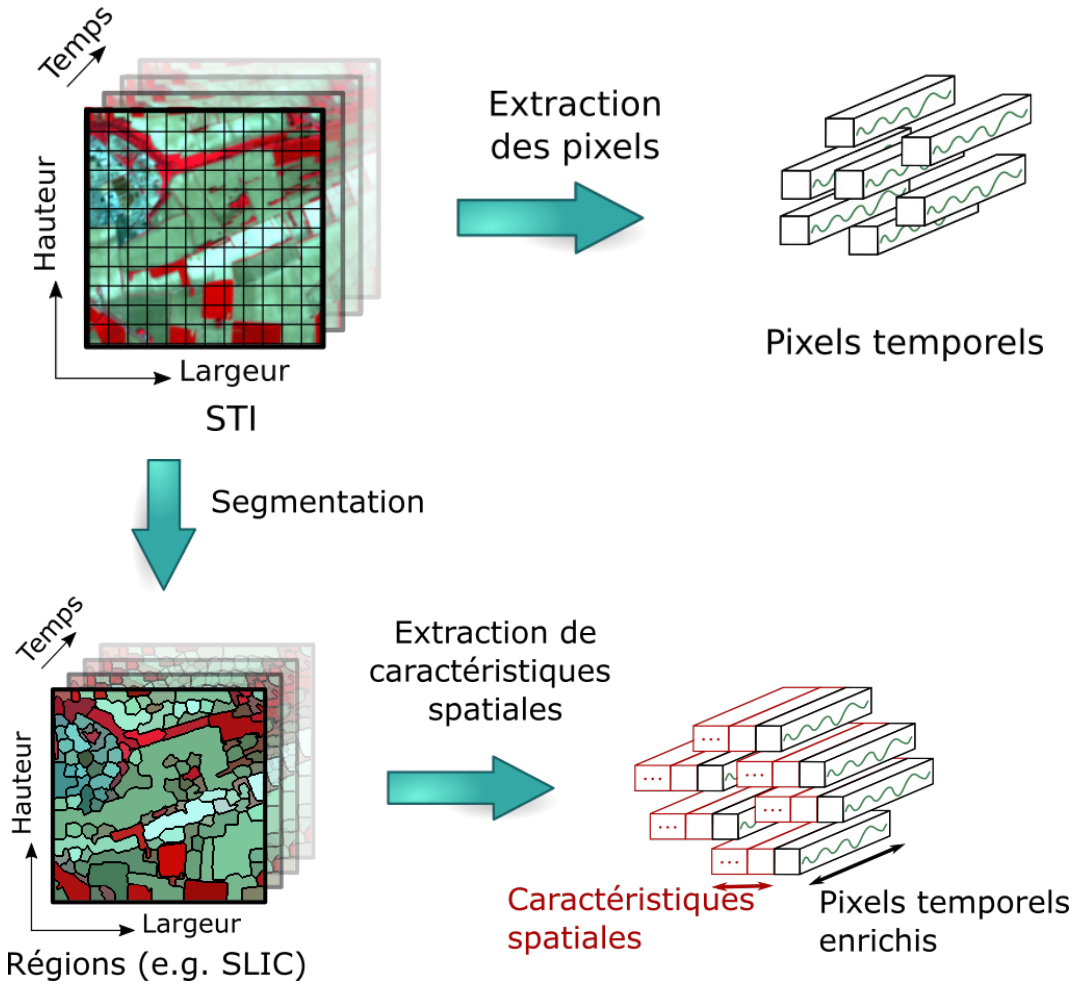
VS.

Apprises
spécifiques

Nature des caractéristiques

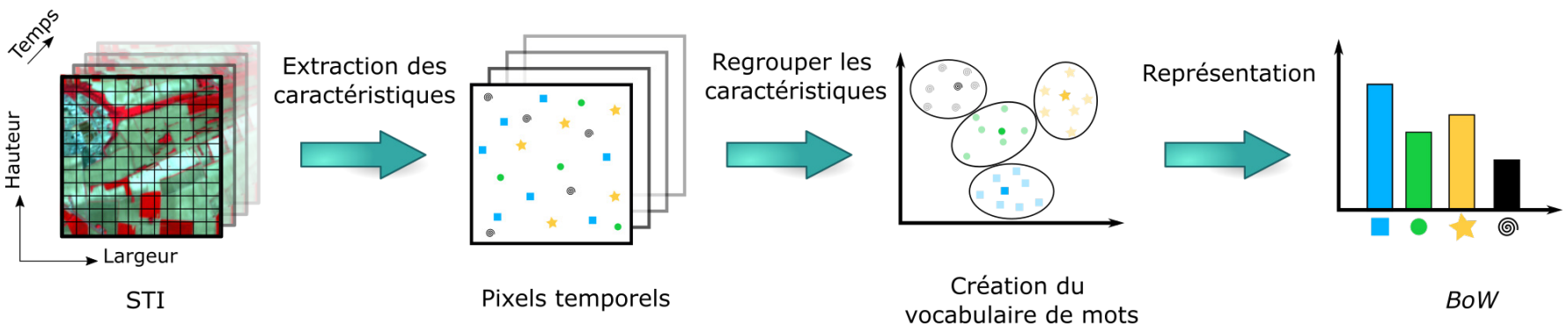


Nature des caractéristiques



Caractéristiques artisanales

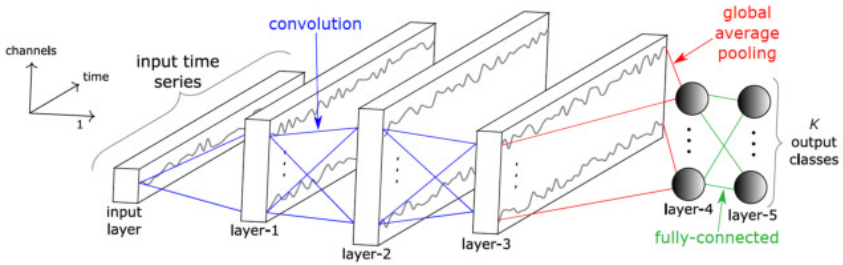
- Exemple de caractéristiques (e.g. HOG [1], SIFT [2])
- Représentation de chaque vidéo par un Sac-de-Mots (noté BoW)
- Classification de chaque vidéo grâce au BoW (SVM ou Forêt Aléatoire)



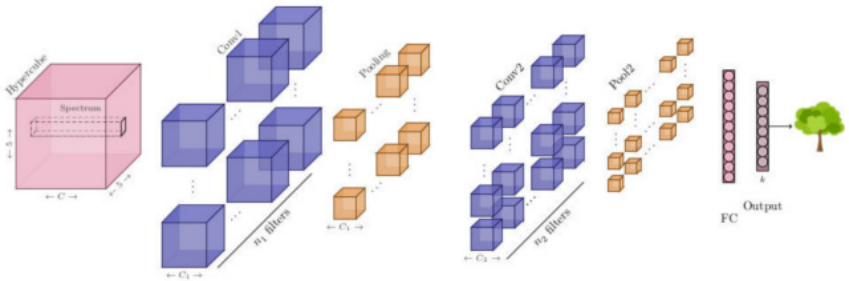
1. Shih-Shinh Huang et al. (2011). "Combining Histograms of Oriented Gradients with Global Feature for Human Detection". In : *Advances in Multimedia Modeling, Part II*. T. 6524. Lecture Notes in Computer Science. Springer, p. 208-218.
2. David G. Lowe (1999). "Object Recognition from Local Scale-Invariant Features". In : *International Conference on Computer Vision*. IEEE Computer Society, p. 1150-1157.

Caractéristiques apprises

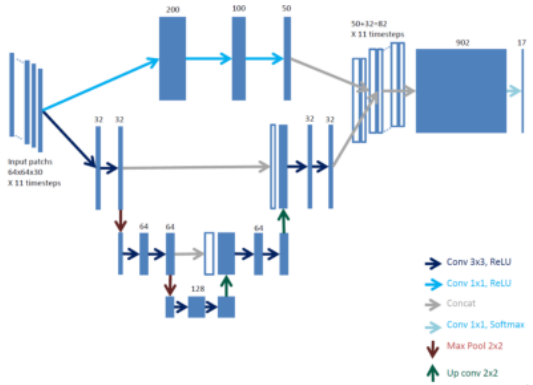
- Analyse des pixels temporels
  - LSTM, TempCNN [1]
- Méthodes hybrides
  - CNN 2D + CNN 1D [2]
- Analyse de séquences temporelles d'images
  - CNN 3D [3]
  - Nuage de points



(a) TempCNN[1]



(b) Convolution 3D[3]



(c) Hybride UNet[2]

1. Charlotte Pelletier, Geoffrey I. Webb et François Petitjean (2019). "Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series". In : *Remote Sensing* 11.5, p. 523.
2. Andrei Stoian et al. (2019). "Land Cover Maps Production with High Resolution Satellite Image Time Series and Convolutional Neural Networks: Adaptations and Limits for Operational Systems". In : *Remote Sensing* 11.17, p. 1986.
3. Nicolas Audebert, Bertrand Le Saux et Sébastien Lefèvre (2019). "Deep Learning for Classification of Hyperspectral Data: A Comparative Review". In : *CoRR* abs/1904.10674.

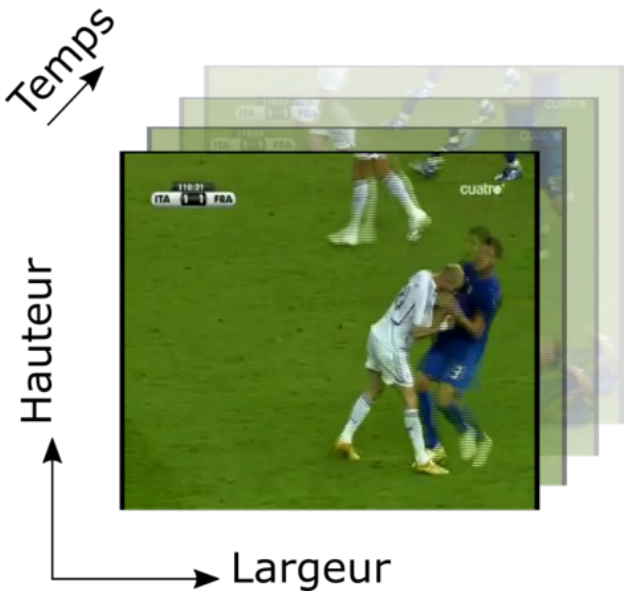
Université de Paris

Classification des séquences temporelles d'images (STI)

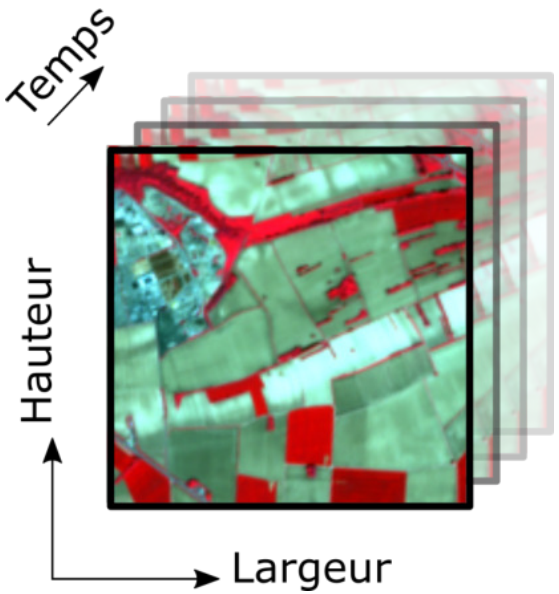
⇒ Compréhension de la dynamique de la scène observée

Challenges et motivations

⇒ Étudier conjointement les domaines spatial et temporel



(a) Avec une continuité visuelle



(b) Avec des images ponctuelles

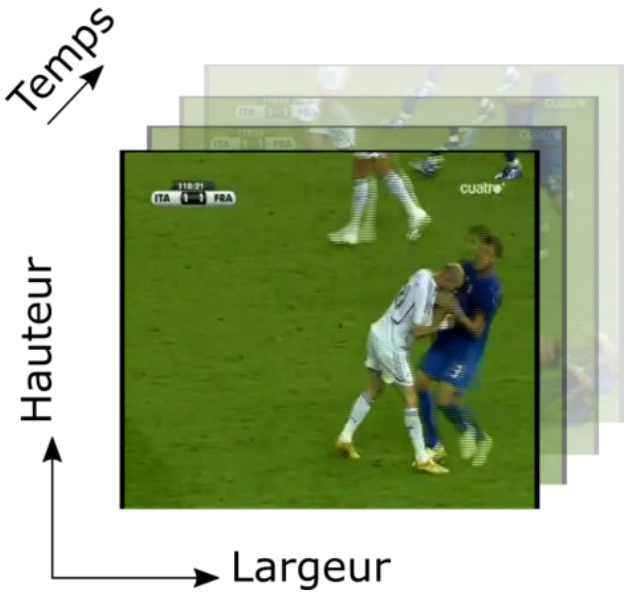
Université de Paris

Classification des séquences temporelles d'images (STI)

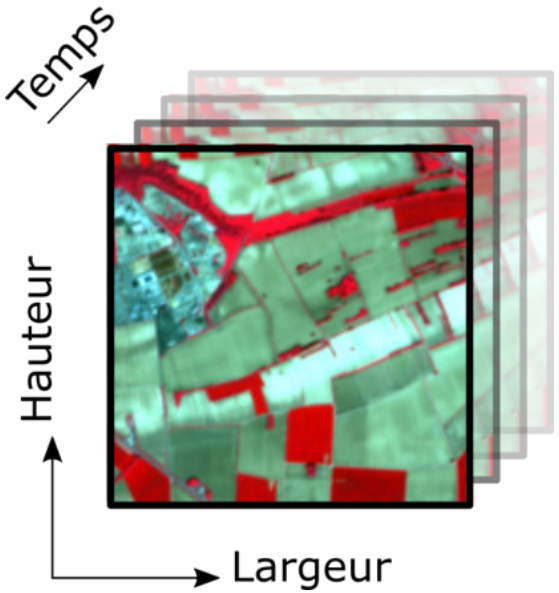
⇒ Compréhension de la dynamique de la scène observée

Challenges et motivations

⇒ Étudier conjointement les domaines spatial et temporel



(a) Avec une continuité visuelle



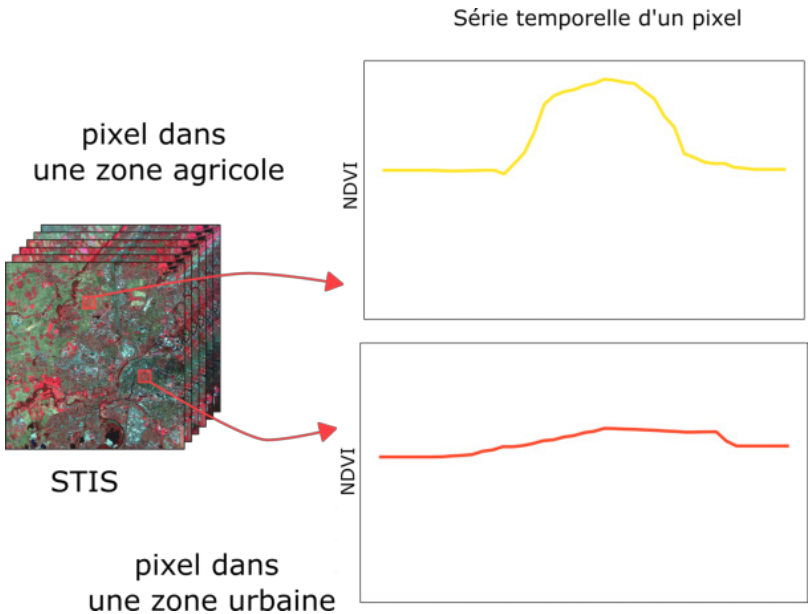
(b) Avec des images ponctuelles

Extraction de caractéristiques spatio-temporelles pour la classification des STI

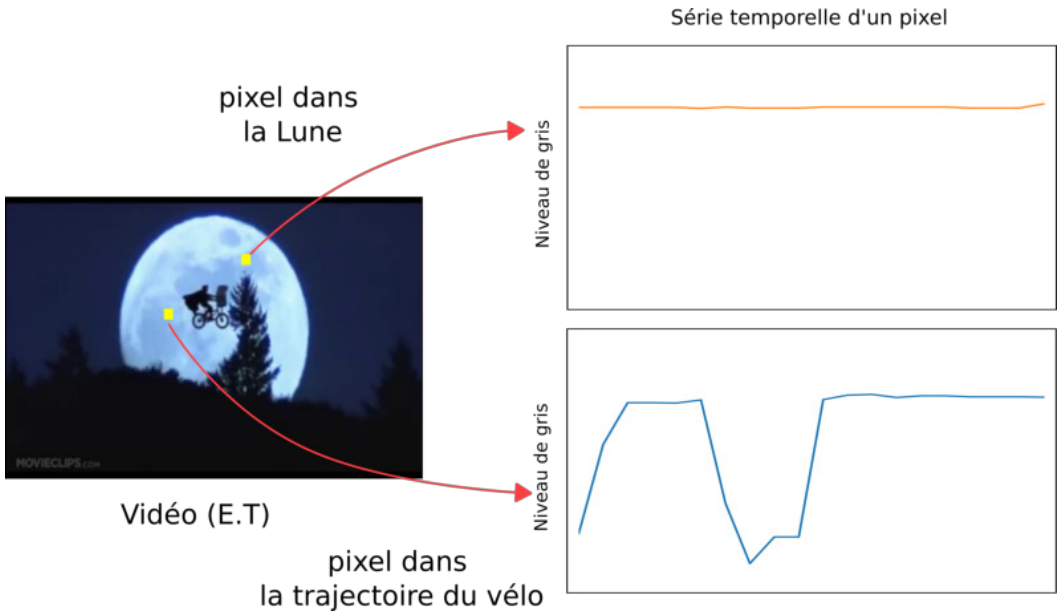
Étude de la stabilité

Rôle de la stabilité

- Étude des zones qui ne subissent pas de changement dans le temps



(a) Application satellite

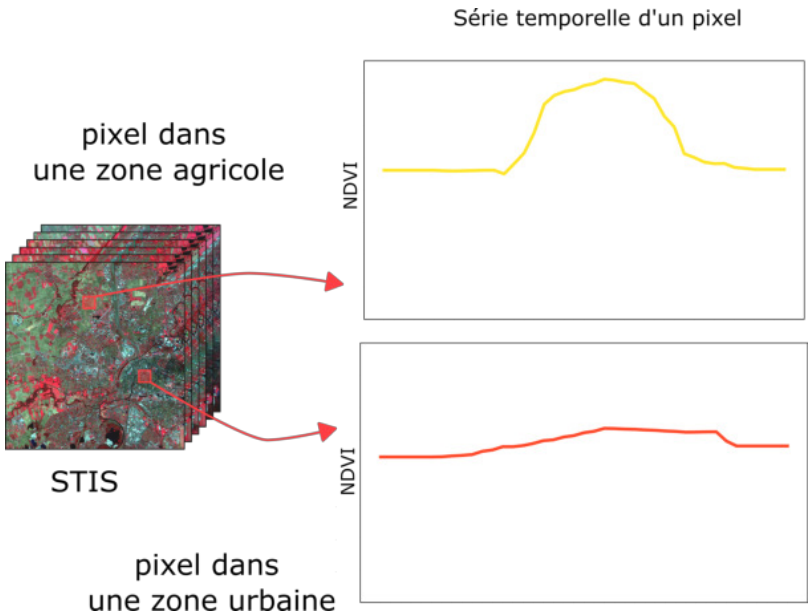


(b) Application vidéo

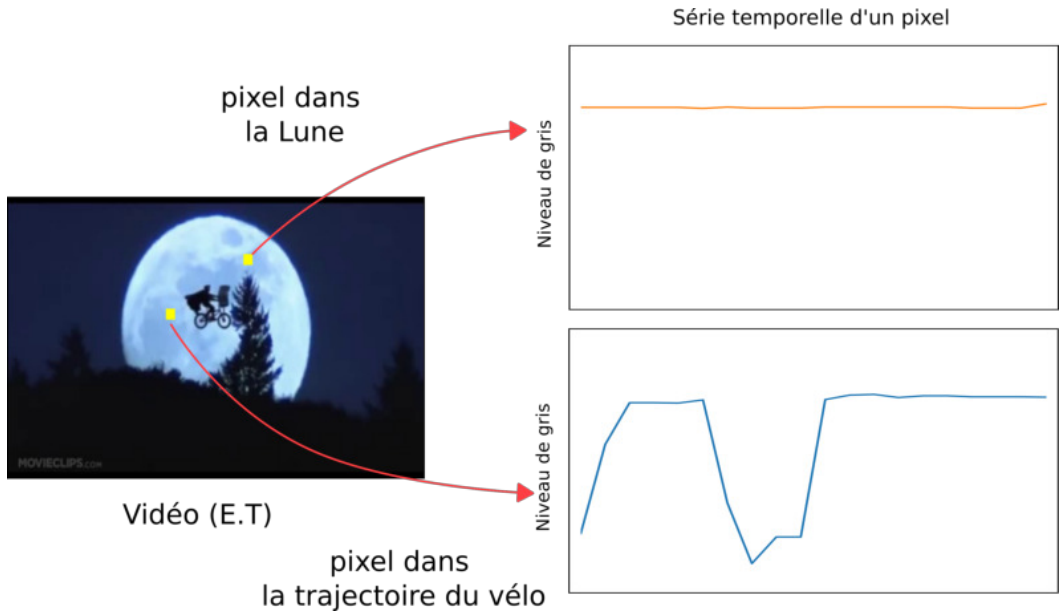
Université de Paris

Rôle de la stabilité

- Étude des zones qui ne subissent pas de changement dans le temps



(a) Application satellite

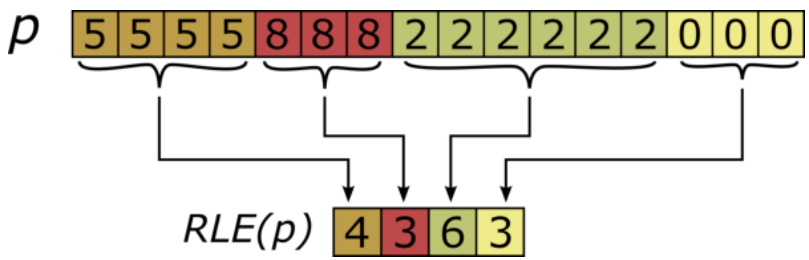


(b) Application vidéo

Au lieu de chercher des changements → nous étudions la stabilité

Vers une nouvelle représentation intermédiaire d'un pixel temporel

- Étudier la répétition des valeurs successives dans le temps
- Quelle stratégie de transformation ?
  - *Run Length Encoding* (RLE) [4]

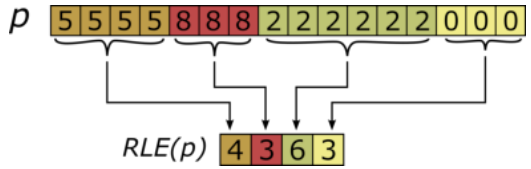


Note : Le RLE est appliqué sur chaque pixel temporel

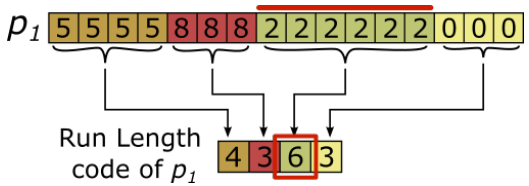
4. Solomon W. Golomb (1966). "Run-length encodings (Corresp.)". In : *IEEE Trans. Inf. Theory* 12.3, p. 399-401.

Université de Paris

Extraction de caractéristiques à partir de la nouvelle représentation

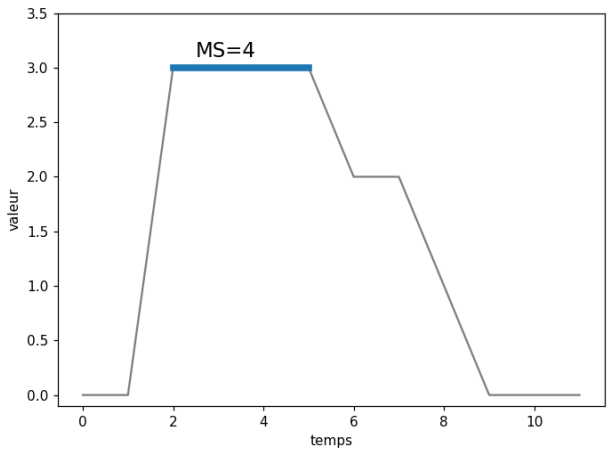


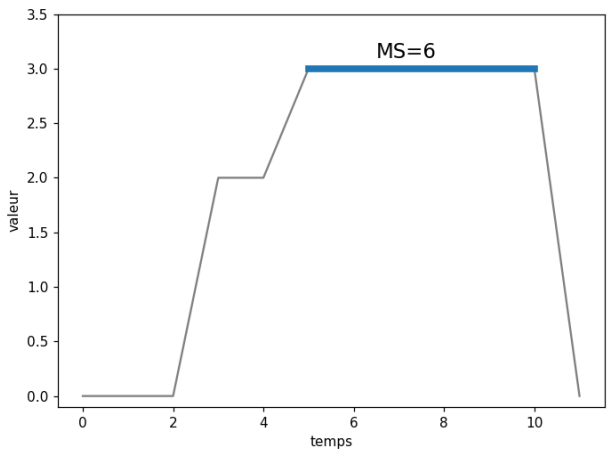
Extraction de caractéristiques à partir de la nouvelle représentation



- **Stabilité maximale** (notée MS)

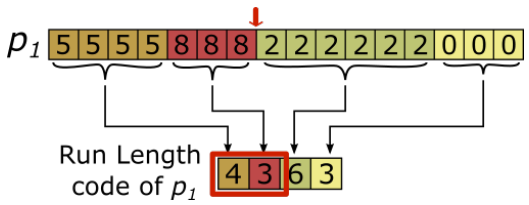
$$MS(p) = \|RLE(p)\|_\infty = \mathbf{6}$$





Université de Paris

Extraction de caractéristiques à partir de la nouvelle représentation



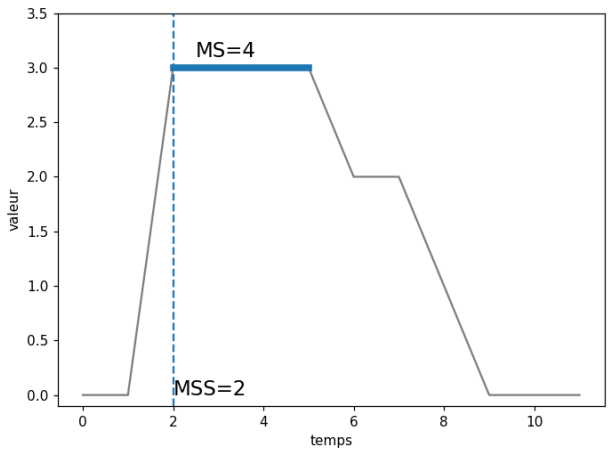
- **Stabilité maximale** (notée *MS*)

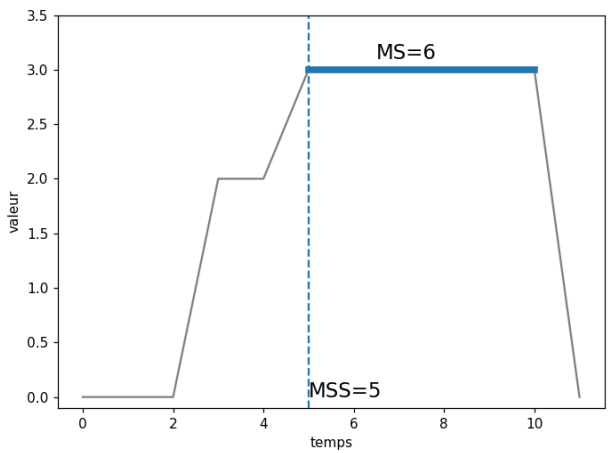
$$MS(p) = \|RLE(p)\|_\infty = \mathbf{6}$$

- **Début de la stabilité maximale** (notée *MSS*)

$$MSS(p) = \left(\sum_{i=1}^{t_0-1} RLE(p)_i\right) = \mathbf{7}$$

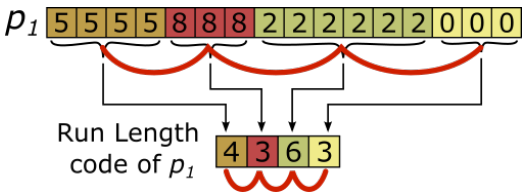
$$\text{avec} \quad t_0 \ / \ RLE(p)_{t_0} = MS(p)$$





Université de Paris

Extraction de caractéristiques à partir de la nouvelle représentation



- **Stabilité maximale** (notée *MS*)

$$MS(p) = \|RLE(p)\|_\infty = \mathbf{6}$$

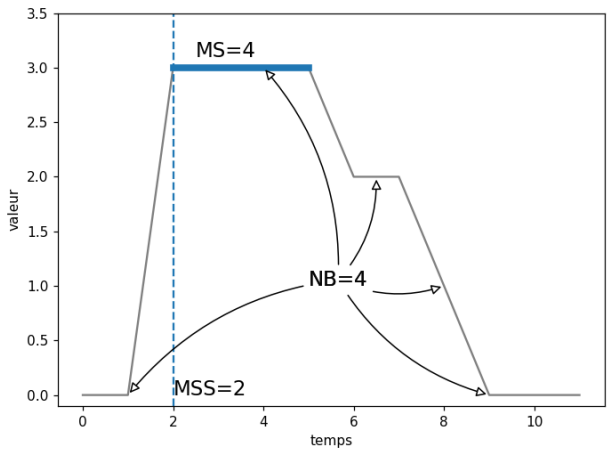
- **Début de la stabilité maximale** (notée *MSS*)

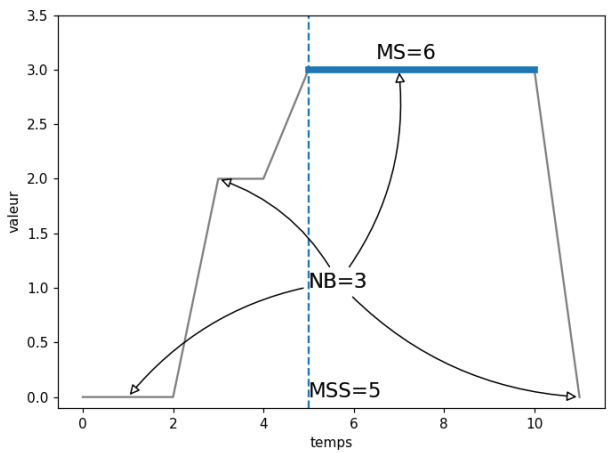
$$MSS(p) = \left(\sum_{i=1}^{t_0-1} RLE(p)_i \right) = \mathbf{7}$$

avec t_0 / $RLE(p)_{t_0} = MS(p)$

- **Nombre de changements** (notée *NB*)

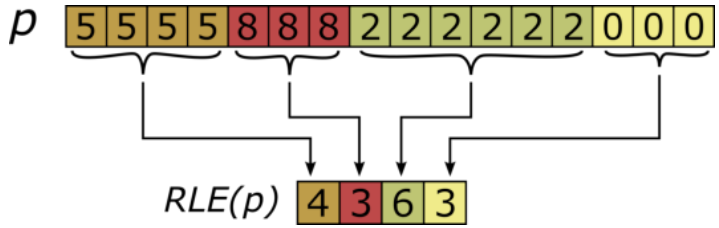
$$NB(p) = l_p - 1 = \mathbf{3}$$



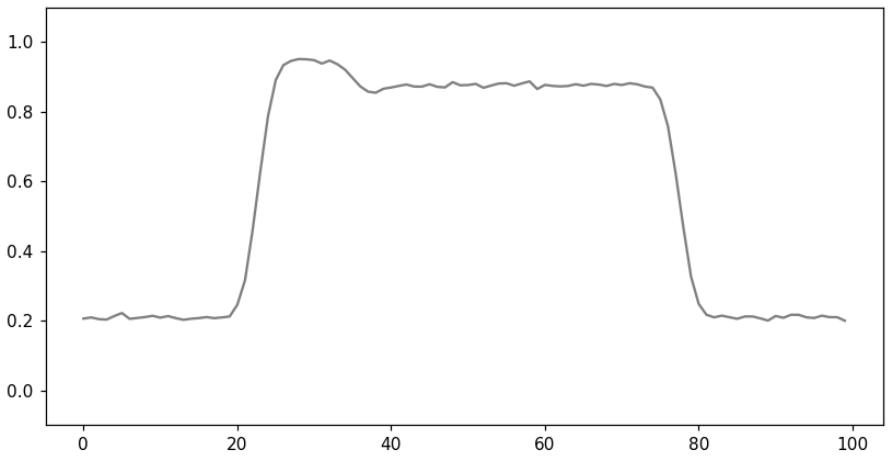


Université de Paris

Extraction de caractéristiques à partir de la nouvelle représentation



La notion d'égalité est la clé pour la mesure de stabilité



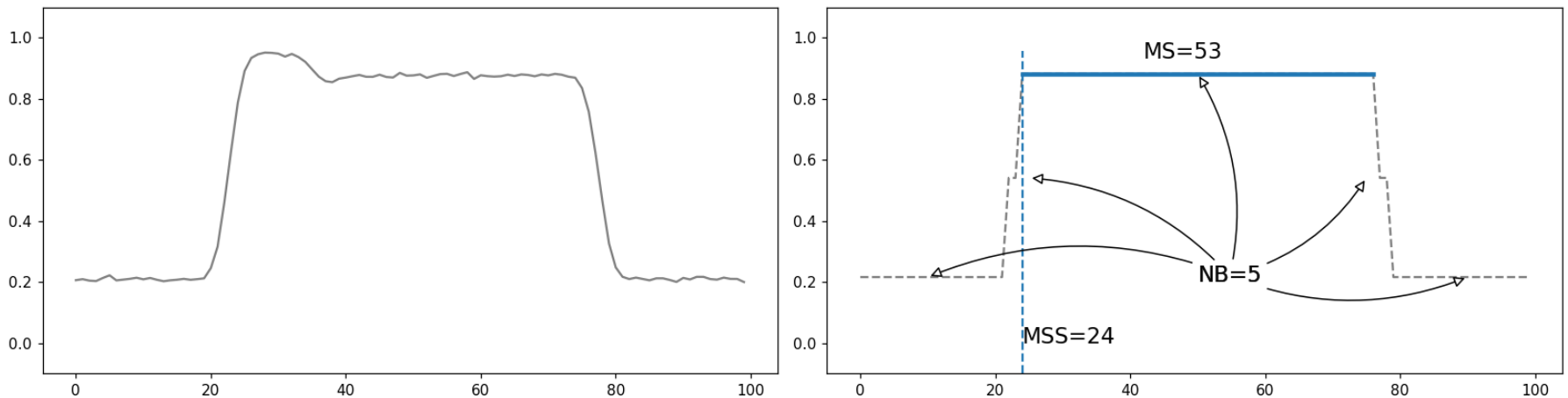
L'égalité des valeurs n'est pas toujours significative

- Variabilité des valeurs dans la séquence temporelle d'images (noté $(I_t)_{t\in[\![1,T]\!]}$)
- Complexité lors de la comparaison vectorielle (e.g. images RGB)

Définition de la notion d'égalité

- Quantification des valeurs de la $(I_t)_{t\in[\![1,T]\!]}$ en appliquant un k-Moyenne (k est un paramètre)
- Nous définissons le prédicat P par l'égalité entre deux objets comme :

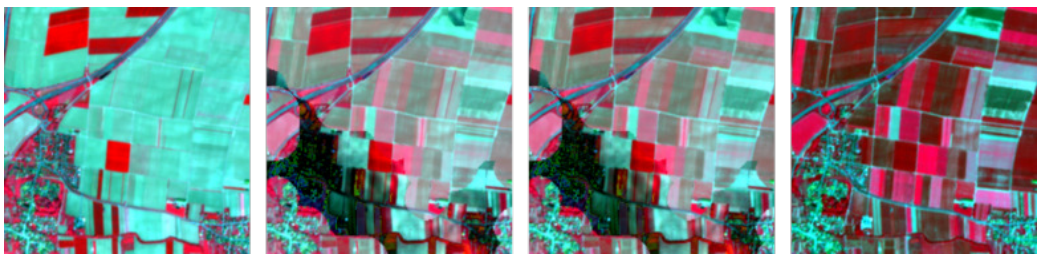
$$P(o_1, o_2) = (o_1^{'} = o_2^{'})$$



Perturbations dans les données

- *Outliers*
- Bruits (e.g. nuages non détectés, artéfacts d'acquisition)
- Alignement des images discrètes

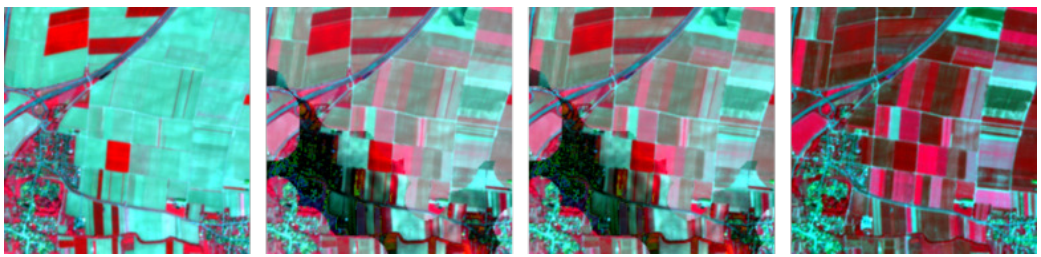
Exemple



Perturbations dans les données

- *Outliers*
- Bruits (e.g. nuages non détectés, artéfacts d'acquisition)
- Alignement des images discrètes

Exemple



Relaxation de l'égalité dans le domaine temporel

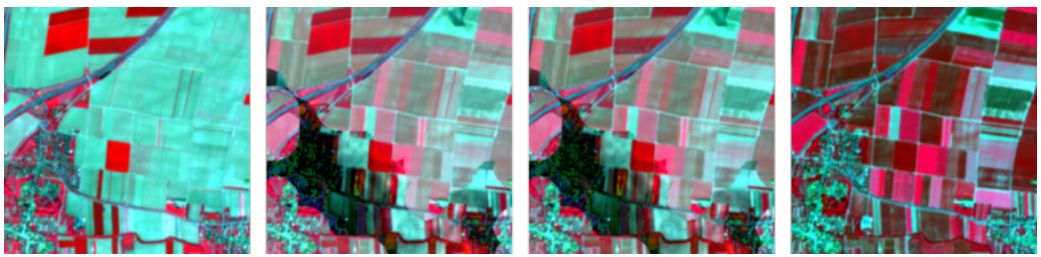
$p(x,y)$	2	5	2	5	2	5	5
$\widetilde{RLE}(p)$			5			2	

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

Perturbations dans les données

- *Outliers*
- Bruits (e.g. nuages non détectés, artéfacts d'acquisition)
- Alignement des images discrètes

Exemple



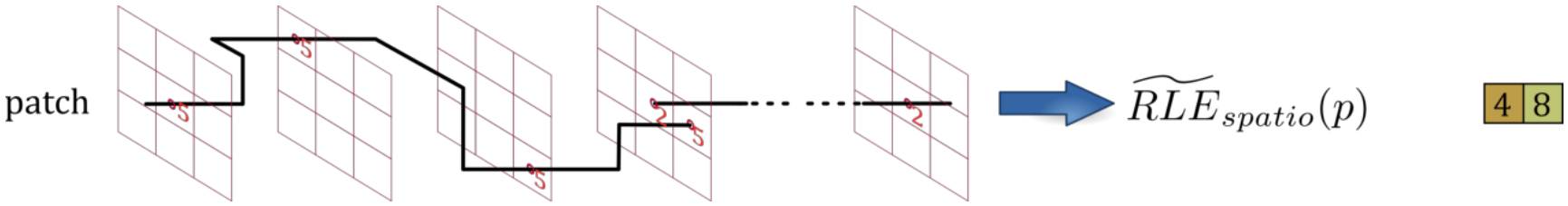
Relaxation de l'égalité dans le domaine temporel

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

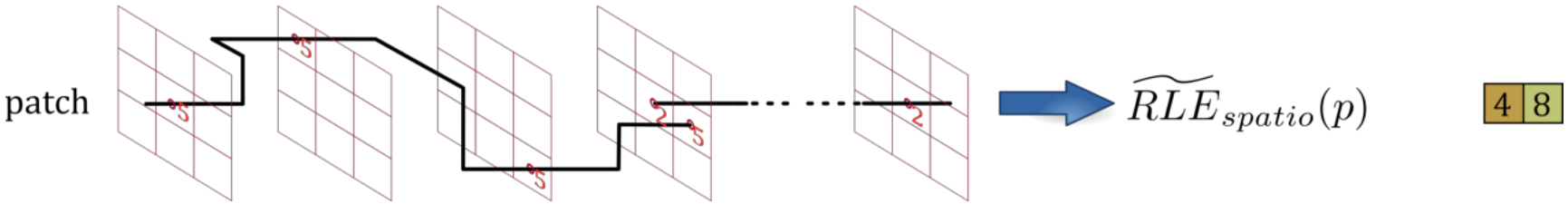
$p(x,y)$	2	5	2	5	2	5	5
$\widetilde{RLE}(p)$	1	6					

Limitée au domaine temporel seulement

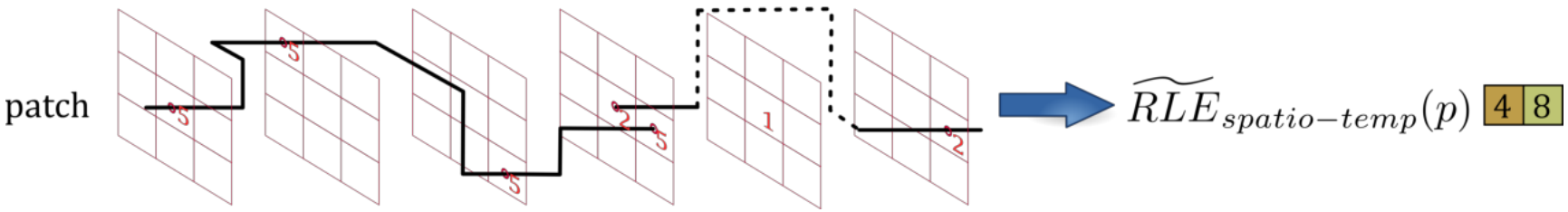
Relaxation de l'égalité dans le domaine spatial



Relaxation de l'égalité dans le domaine spatial



Relaxation de l'égalité dans le domaine spatio-temporel



Composition des caractéristiques extraites en une image en fausses couleurs

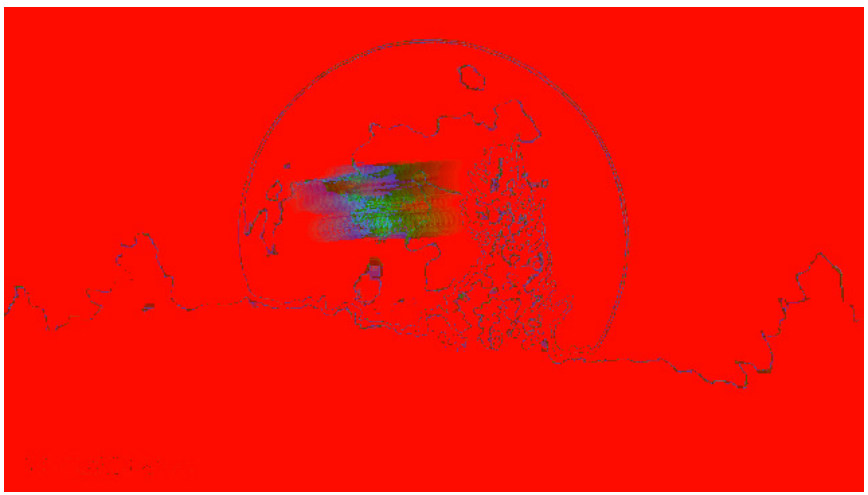
- Stabilité Max
- Nb changements
- Début de la Stabilité Max

Vidéo originale

Composition des caractéristiques extraites en une image en fausses couleurs

- Stabilité Max
- Nb changements
- Début de la Stabilité Max

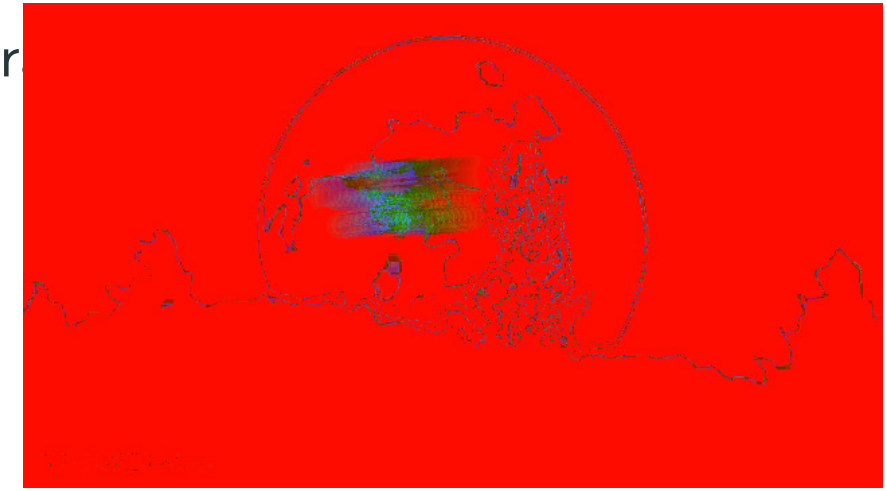
Vidéo originale



Résumé *TS*

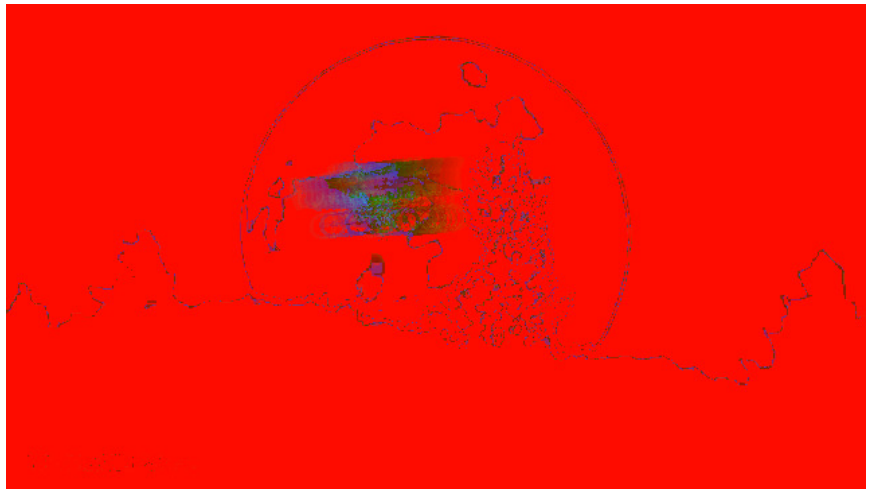
Composition des caractéristiques extr
couleurs

- Stabilité Max
- Nb changements
- Début de la Stabilité Max



Résumé *TS*

Vidéo originale

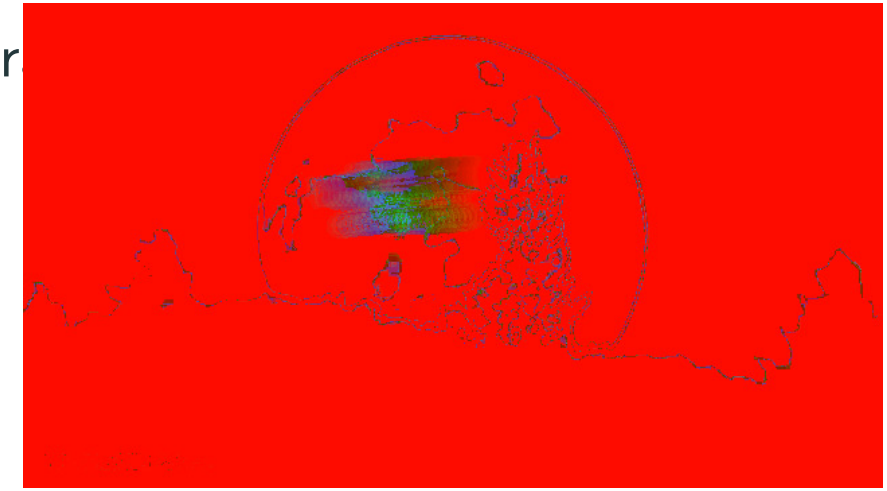


Résumé TS_{temp}

Université de Paris

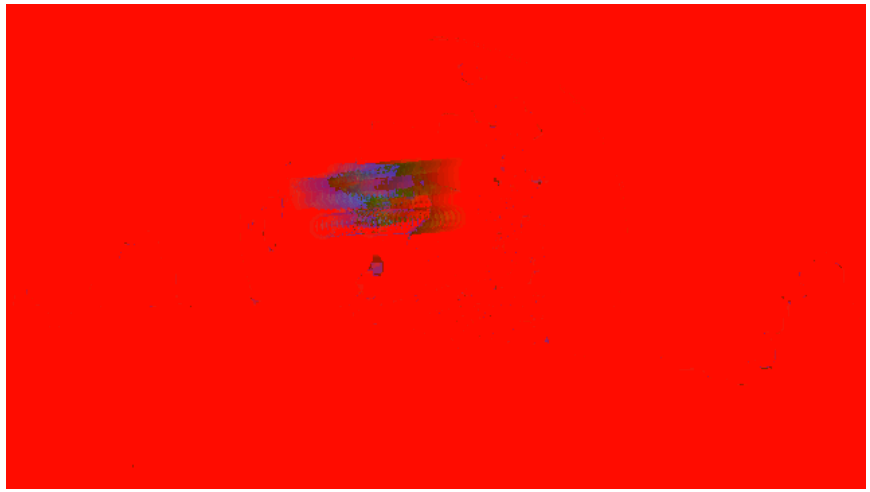
Composition des caractéristiques extr couleurs

- Stabilité Max
- Nb changements
- Début de la Stabilité Max



Résumé *TS*

Vidéo originale

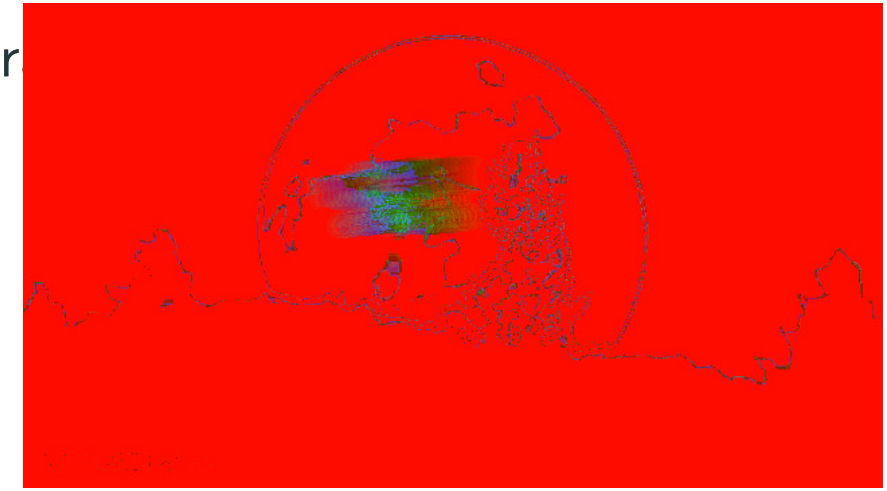


Résumé TS_{spatio}

Université de Paris

Composition des caractéristiques extr
couleurs

- Stabilité Max
- Nb changements
- Début de la Stabilité Max



Résumé *TS*

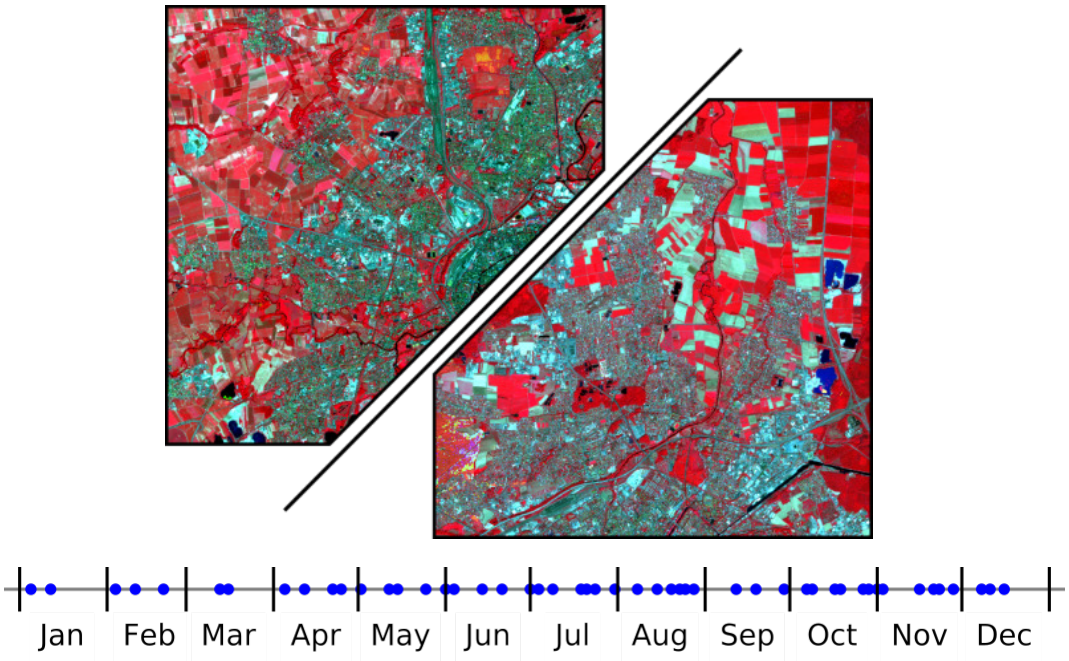
Vidéo originale

Résumé $TS_{spatio-temp}$

Application et évaluation sur deux cadres applicatifs :

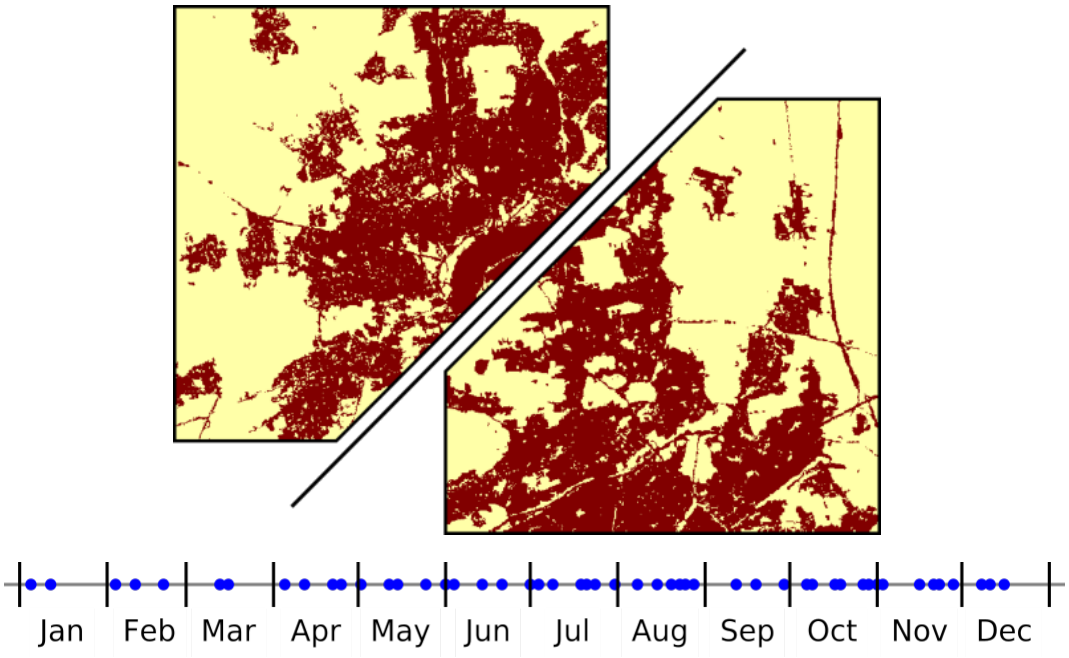
- Application de télédétection
  Analyse de la tache urbaine
- Application de vidéo
  Reconnaissance de scènes de violence

Données pour l'analyse de la couverture urbaine



- Série Temporelle d'Images Satellitaires issue de Sentinel-2 de 2017
  - 50 images de taille 1000×1000
  - Chaque pixel est caractérisé par l'indice de végétation : $NDVI = \frac{Nir-R}{Nir+R}$
- Interpolation linéaire sur les zones masquées par des nuages
- Deux villes sont sélectionnées : **Strasbourg** et **Mulhouse**

Données pour l'analyse de la couverture urbaine



- Données de référence
  - **Produit d'imperméabilité des matériaux** : représente le pourcentage d'imperméabilisation du sol
  - **Classes** : zones artificielles ($>$ 0% imperméabilité)
    zones naturelles (0% imperméabilité)

Classification des STIS

Caractéristiques :

- Pixel temporel p^{NDVI} (50 carac.)
- Moyenne du pixel temporel $\overline{p^{NDVI}}$ (1 carac.)
- TS, TS_{temp}, TS_{spatio} et $TS_{spatio-temp}$ (3 carac.)

Classificateurs :

- Arbre de décision : nb carac. $\leqslant 3$
- Forêt aléatoire : nb carac. > 3
- TempCNN [5] : nb carac. > 3

5. Charlotte Pelletier, Geoffrey I. Webb et François Petitjean (2019). "Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series". In : *Remote Sensing* 11.5, p. 523.

Classification des STIS

Caractéristiques :

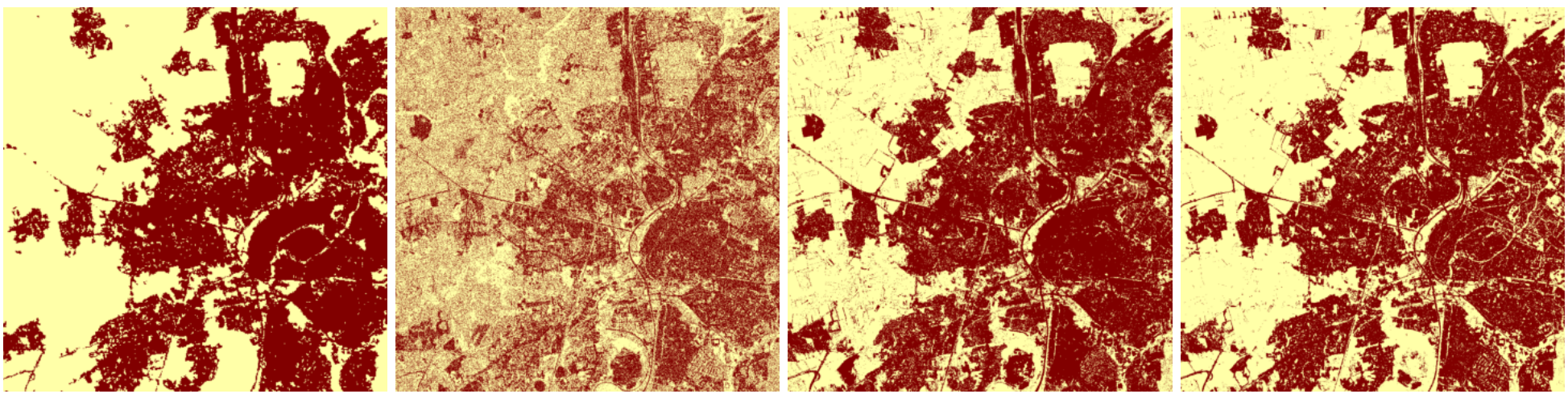
- Pixel temporel p^{NDVI} (50 carac.)
- Moyenne du pixel temporel $\overline{p^{NDVI}}$ (1 carac.)
- TS, TS_{temp}, TS_{spatio} et $TS_{spatio-temp}$ (3 carac.)

Classificateurs :

- Arbre de décision : nb carac. $\leqslant 3$
- Forêt aléatoire : nb carac. > 3
- TempCNN [5] : nb carac. > 3

Expérience :

- Entraînement sur **Mulhouse** et test sur **Strasbourg**



Référence | $\overline{p^{NDVI}}$ | $TS_{spatio-temp}$ | $p^{NDVI} + TS_{spatio-temp}$

5. Charlotte Pelletier, Geoffrey I. Webb et François Petitjean (2019). "Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series". In : *Remote Sensing* 11.5, p. 523.

Classification des STIS

Caractéristiques :

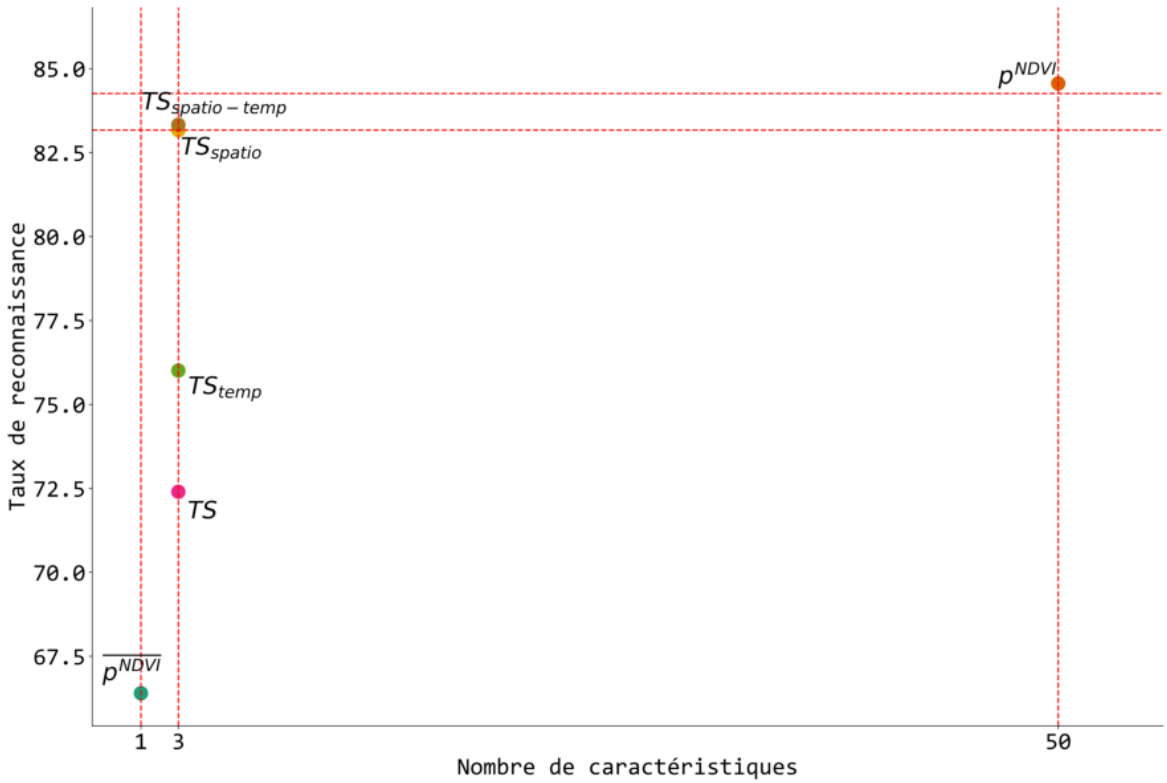
- Pixel temporel p^{NDVI} (50 carac.)
- Moyenne du pixel temporel $\overline{p^{NDVI}}$ (1 carac.)
- TS, TS_{temp}, TS_{spatio} et $TS_{spatio-temp}$ (3 carac.)

Classificateurs :

- Arbre de décision : nb carac. $\leqslant 3$
- Forêt aléatoire : nb carac. > 3
- TempCNN [5] : nb carac. > 3

Expérience :

- Entraînement sur **Mulhouse** et test sur **Strasbourg**



5. Charlotte Pe… …Network for the Classificati…

Classification des STIS

Caractéristiques :

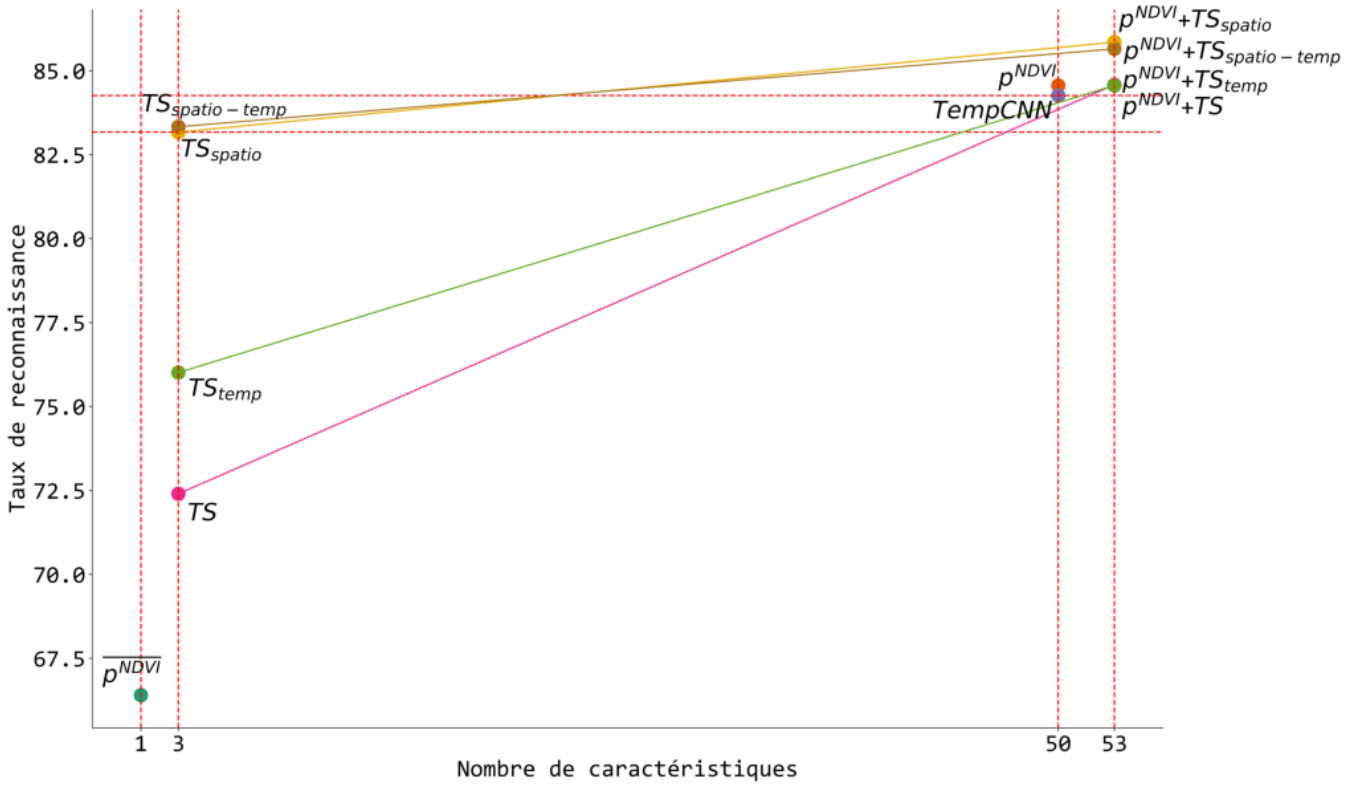
- Pixel temporel p^{NDVI} (50 carac.)
- Moyenne du pixel temporel $\overline{p^{NDVI}}$ (1 carac.)
- TS, TS_{temp}, TS_{spatio} et $TS_{spatio-temp}$ (3 carac.)

Classificateurs :

- Arbre de décision : nb carac. $\leqslant 3$
- Forêt aléatoire : nb carac. > 3
- TempCNN [5] : nb carac. > 3

Expérience :

- Entraînement sur **Mulhouse** et test sur **Strasbourg**



5. Charlotte Pe… …Network for the Classificati…

Application et évaluation sur deux cadres applicatifs :

- Application de télédétection
  Analyse de la tache urbaine
- Application de vidéo
  Reconnaissance de scènes de violence

Université de Paris

Données et vérité terrain pour la reconnaissance de scènes de violence

(a) RWF2000

(b) Movies Fights

(c) Hockey Fights

(d) Crowd Violence

- Quatre jeux de données sont utilisés
- Nombre de vidéos par classe équilibré dans tous les datasets

Dataset	**RWF2000**[1]	**Movies fights**[2]	**Hockey fights**[2]	**Crowd Violence**[3]
Nb de vidéos	2000	200	1000	246
		Propriétés		
Durée (second)	150	42-60	40-49	26-163
Largeur (pixel)	204-1920	720	360	320
Hauteur (pixel)	188 - 1080	480-576	288	240

1. M. Cheng, and al Rwf-2000 : An open large scale video database for violence detection, in ICPR, 2020, pp. 4183–4190.
2. E. B. Nievas, and al., Violence detection in video using computer vision techniques, in CAIP.
3. Y. I. T. Hassner and O. Kliper-Gross, Violent flows : Real-time detection of violent crowd behavior, in CVPR workshops, 2012, pp. 1–6.

Visualisation du résumé 2D

- Utilisation de vidéos en niveaux de gris
- Quantification des valeurs avec k-Moyenne : $k = 4$

$\Rightarrow$ Composition des résumés TS_*
- Stabilité Max
- Nb changements
- Début de la Stab. Max

Vidéo violente

TS

TS_{temp}

TS_{spatio}

$TS_{spatio-temp}$

Université de Paris

Classification des vidéos

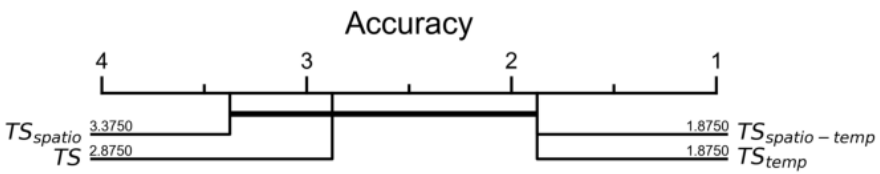
- Modèle : SqueezeNet (Iandola et al., 2016)

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**

Classification des vidéos

- Modèle : SqueezeNet (Iandola et al., 2016)

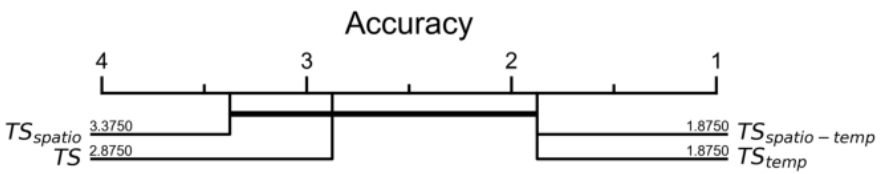
k_{quanti}	**Caractéristique**	*RWF2000*	*Movies fights*	*Hockey fights*	*Crowd Violence*
4	TS	82.5	**97.5**	88.6	80.0
	TS_{temp}	**82.7**	**97.5**	89.9	83.3
	TS_{spatio}	81.5	**97.5**	88.2	80.8
	$TS_{spatio-temp}$	81.7	**97.5**	**91.1**	**84.5**



Classification des vidéos

- Modèle : SqueezeNet (Iandola et al., 2016)

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**

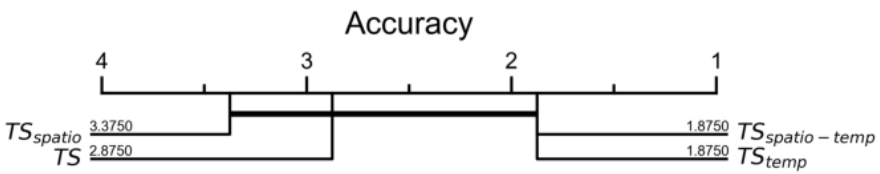


Méthode	*RWF2000*	*Movies fights*	*Hockey fights*	*Crowd Violence*
CNN 3D				
Temp. Seg. Nets (Wang et al., 2016)	81.5	94.2	91.5	81.5
I3D (Carreira et Zisserman, 2017)	83.4	95.8	93.4	83.4
Represent. flow (Wang et al., 2017)	85.3	**97.3**	92.5	**85.9**
Flow Gated Net (Cheng, Cai et Li, 2020)	**87.3**	n/a	**98.0**	88.8
ECO (Zolfaghari, Singh et Brox, 2018)	83.7	96.3	94.0	84.7
Nuage de points				
PointNet++ (Qi et al., 2017)	78.2	89.2	89.7	89.2
PointConv (Wu, Qi et Li, 2019)	76.8	91.3	89.7	89.2
DGCNN (Wang et al., 2019)	80.6	92.6	90.2	87.4
SPIL (Su et al., 2020)	**89.3**	**98.5**	**96.8**	**94.5**

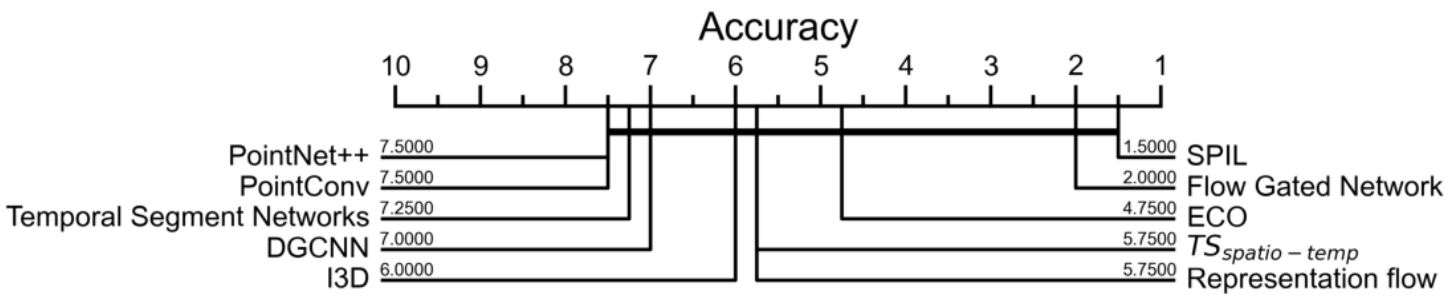
Classification des vidéos

- Modèle : SqueezeNet (Iandola et al., 2016)

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**



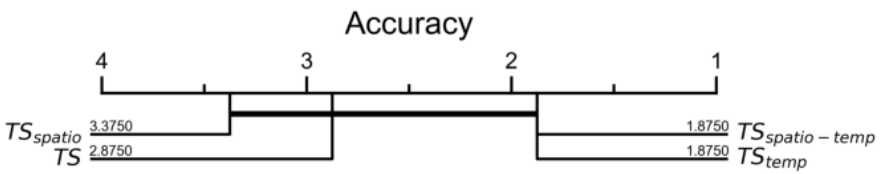
Méthode	*RWF2000*	*Movies fights*	*Hockey fights*	*Crowd Violence*
CNN 3D				
Temp. Seg. Nets (Wang et al., 2016)	81.5	94.2	91.5	81.5
I3D (Carreira et Zisserman, 2017)	83.4	95.8	93.4	83.4
Represent. flow (Wang et al., 2017)	85.3	**97.3**	92.5	**85.9**
Flow Gated Net (Cheng, Cai et Li, 2020)	**87.3**	n/a	**98.0**	88.8
ECO (Zolfaghari, Singh et Brox, 2018)	83.7	96.3	94.0	84.7
Nuage de points				
PointNet++ (Qi et al., 2017)	78.2	89.2	89.7	89.2
PointConv (Wu, Qi et Li, 2019)	76.8	91.3	89.7	89.2
DGCNN (Wang et al., 2019)	80.6	92.6	90.2	87.4
SPIL (Su et al., 2020)	**89.3**	**98.5**	**96.8**	**94.5**



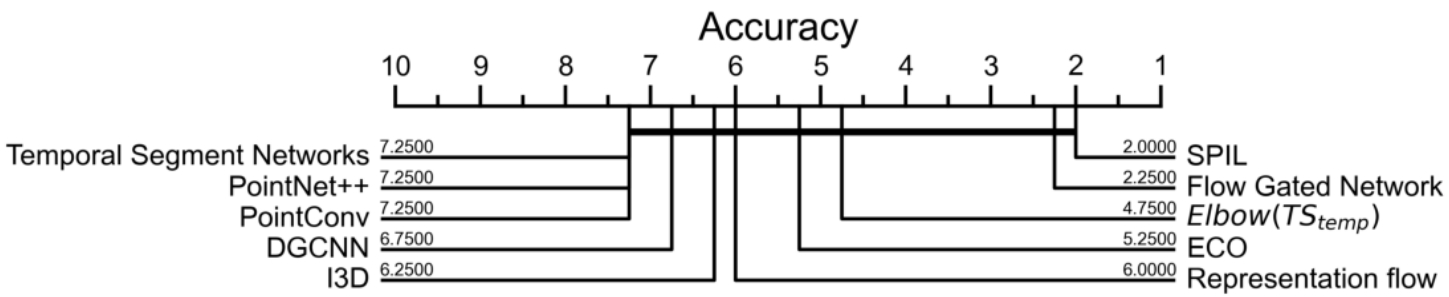
Classification des vidéos

- Modèle : SqueezeNet (Iandola et al., 2016)

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**



Méthode	*RWF2000*	*Movies fights*	*Hockey fights*	*Crowd Violence*
CNN 3D				
Temp. Seg. Nets (Wang et al., 2016)	81.5	94.2	91.5	81.5
I3D (Carreira et Zisserman, 2017)	83.4	95.8	93.4	83.4
Represent. flow (Wang et al., 2017)	85.3	**97.3**	92.5	**85.9**
Flow Gated Net (Cheng, Cai et Li, 2020)	**87.3**	n/a	**98.0**	88.8
ECO (Zolfaghari, Singh et Brox, 2018)	83.7	96.3	94.0	84.7
Nuage de points				
PointNet++ (Qi et al., 2017)	78.2	89.2	89.7	89.2
PointConv (Wu, Qi et Li, 2019)	76.8	91.3	89.7	89.2
DGCNN (Wang et al., 2019)	80.6	92.6	90.2	87.4
SPIL (Su et al., 2020)	**89.3**	**98.5**	**96.8**	**94.5**



Ce qu'il faut retenir

- Extraction de caractéristiques artisanales qui mesurent la stabilité temporelle
- Relaxation de l'égalité pour rendre ces caractéristiques spatio-temporelles
- Utilisation :
  - Résumer une séquence temporelle d'images
  - Classifier les données

Avantages et limites

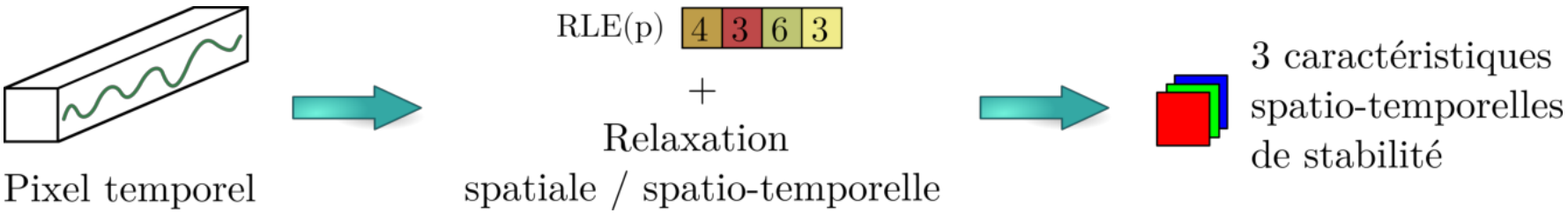
- Facilité d'analyse des territoires urbains avec les données satellitaires
- Présente des limites en analyse de vidéos
  - Déplacement des objets ou de la caméra

Code source : `https://github.com/mchelali/TemporalStability`

Étude des variations des séquences temporelles d'images

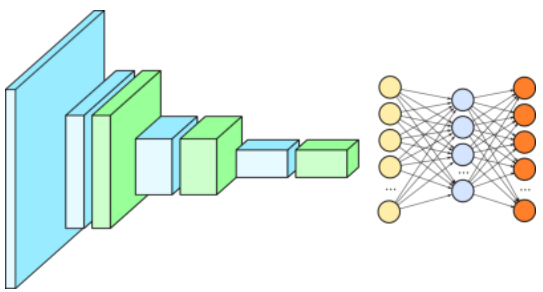
Université de Paris

Pourquoi les variations ?



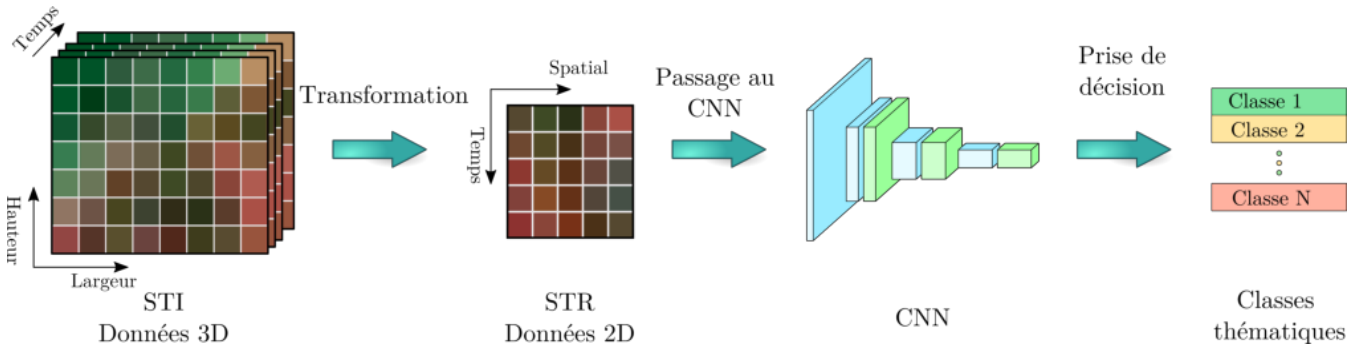
Information spatiale très limitée

Utilisation des réseaux de neurones convolutionnels

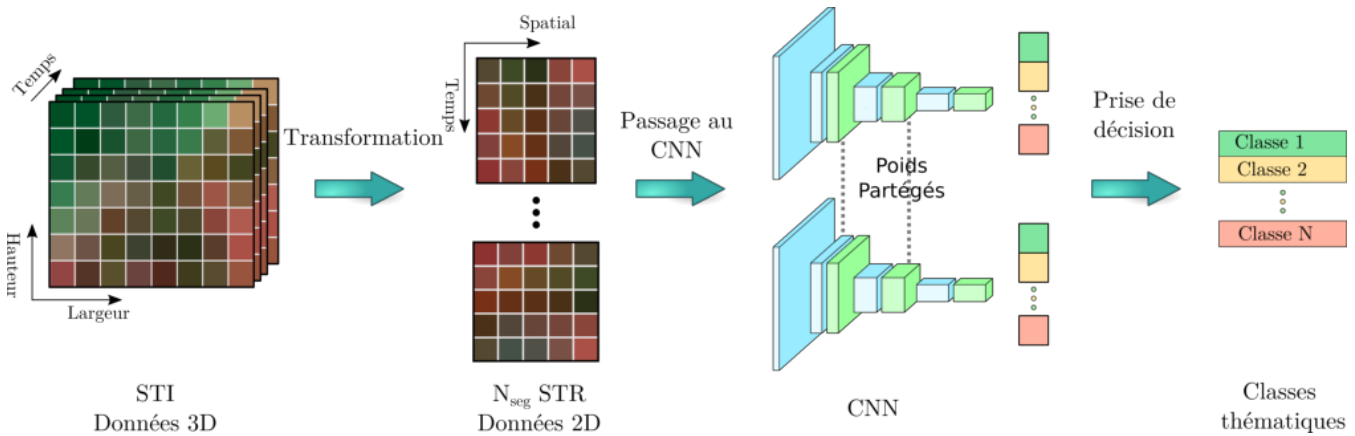


Deep-STaR : composée de deux étapes

- Hors-ligne

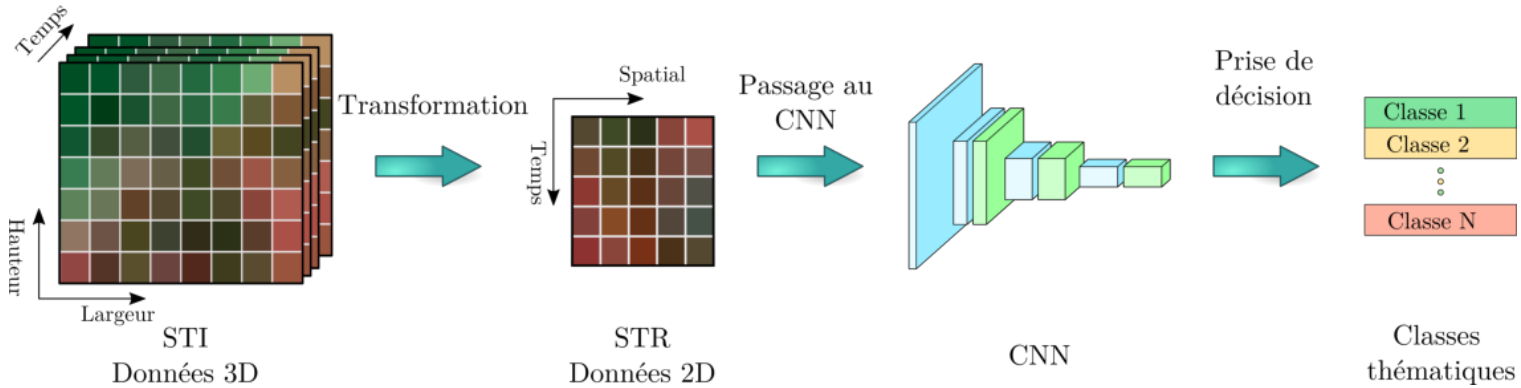


- En-ligne

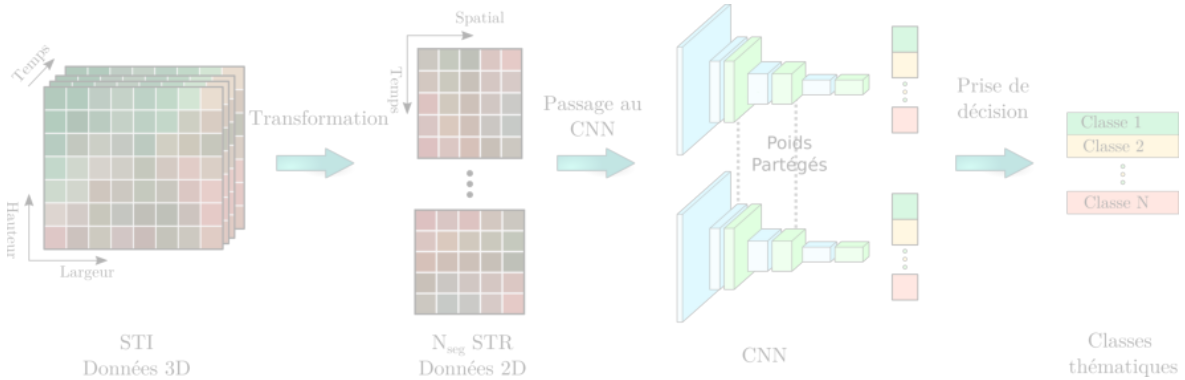


DEEP-STAR : composée de deux étapes

- Hors-ligne



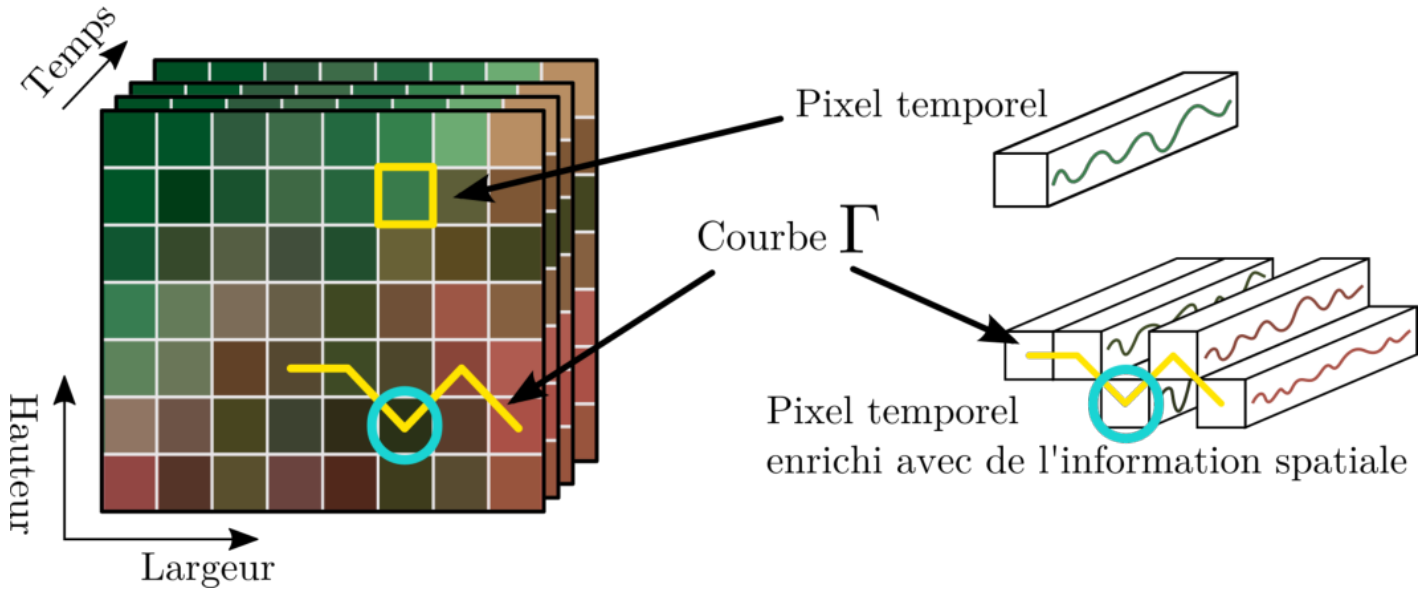
- En-ligne



Deep-STaR : du 3D au STR

Stratégie :

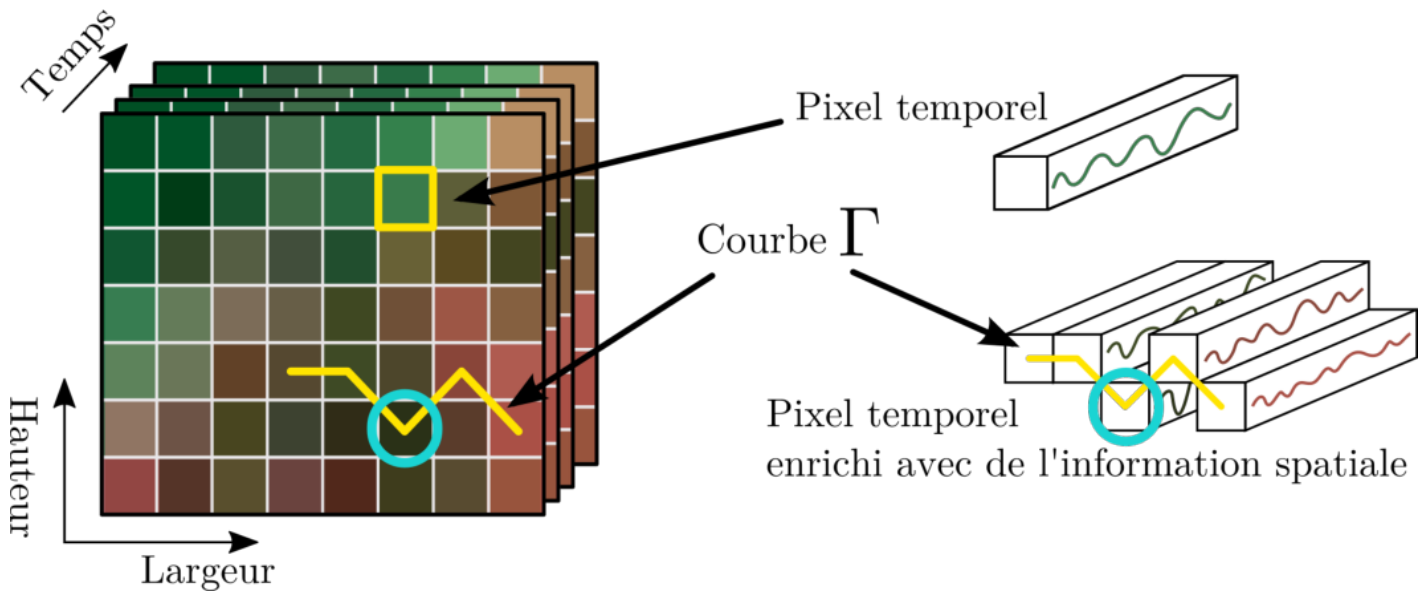
- Observer la STI d'un point de vue intermédiaire



Deep-STaR : du 3D au STR

Stratégie :

- Observer la STI d'un point de vue intermédiaire
- Représenter une STI par plusieurs représentations spatio-temporelles (STR)



⇒ Permet de traiter des images 2D (au lieu du cube 3D)

Université de Paris

Représentation spatio-temporelle des séquences temporelles d'images

- Réduire la complexité de la structure des données : 2D+t à 2D
- **Transformation de chaque image en un vecteur 1D**

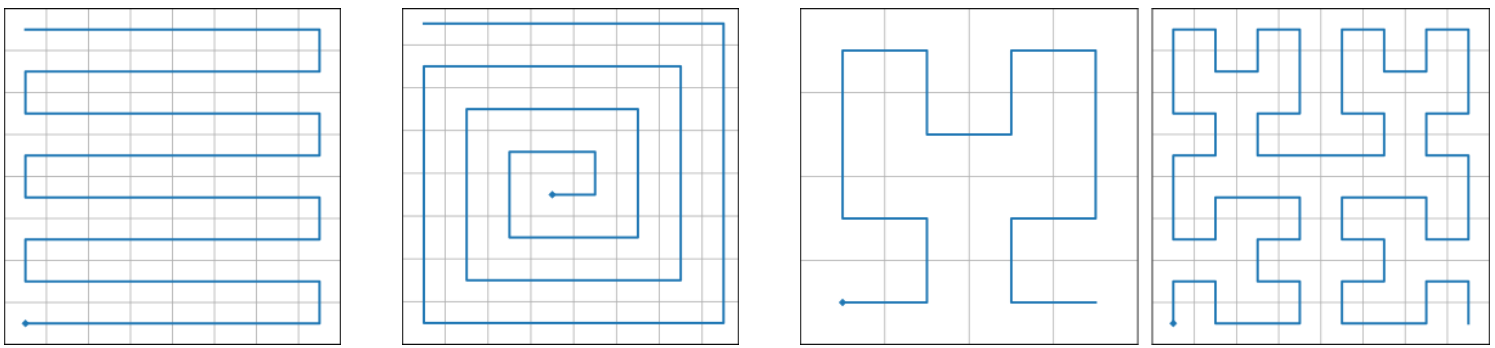
Problème : Perte partielle de l'information spatiale

Représentation spatio-temporelle des séquences temporelles d'images

- Réduire la complexité de la structure des données : 2D+t à 2D
- **Transformation de chaque image en un vecteur 1D**

Problème : Perte partielle de l'information spatiale

$\Rightarrow$ Courbes globales ou courbes remplissant l'espace



(a) $\Re_{snake}$ **(b)** $\Re_{spiral}$ **(c)** $\Re_{Hilbert}$ (2^{eme} et 3^{eme} ordres)

Université de Paris

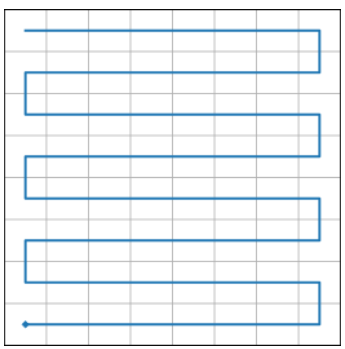
Représentation spatio-temporelle des séquences temporelles d'images

- Réduire la complexité de la structure des données : 2D+t à 2D
- **Transformation de chaque image en un vecteur 1D**

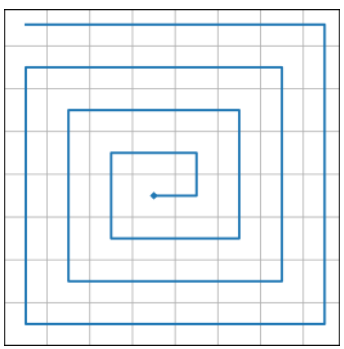
Problème : Perte partielle de l'information spatiale

$\Rightarrow$ Courbes globales ou courbes remplissant l'espace

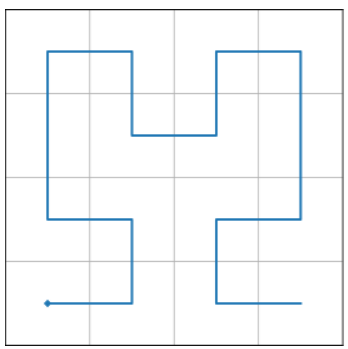
$\Rightarrow$ Courbes locales : *Random Walk (RW)*



(a) $\Re_{snake}$



(b) $\Re_{spiral}$



(c) $\Re_{Hilbert}$ (2^{eme} et 3^{eme} ordres)

(d) *RW*

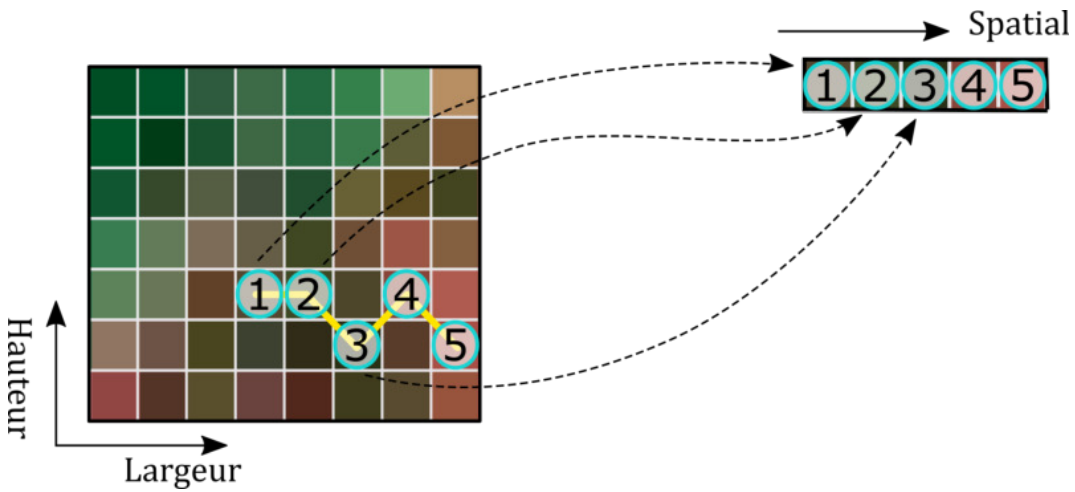
Représentation spatio-temporelle des séquences temporelles d'images

- Réduire la complexité de la structure des données : 2D+t à 2D
- **Transformation de chaque image en un vecteur 1D**

Problème : Perte partielle de l'information spatiale

⇒ Courbes globales ou courbes remplissant l'espace

⇒ Courbes locales : *Random Walk (RW)*



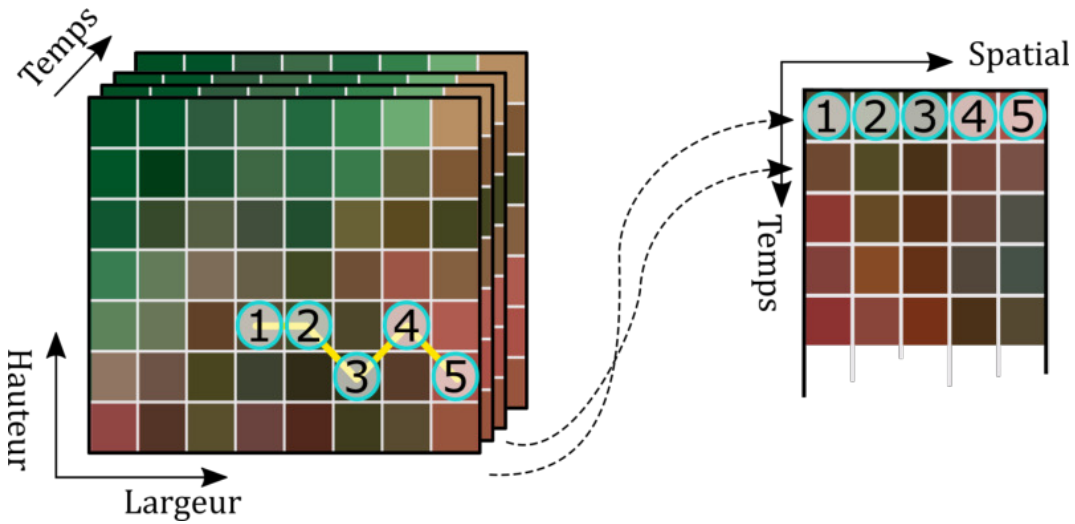
Représentation spatio-temporelle des séquences temporelles d'images

- Réduire la complexité de la structure des données : 2D+t à 2D
- **Transformation de chaque image en un vecteur 1D**

Problème : Perte partielle de l'information spatiale

⇒ Courbes globales ou courbes remplissant l'espace

⇒ Courbes locales : *Random Walk (RW)*



Représentation spatio-temporelle des séquences temporelles d'images

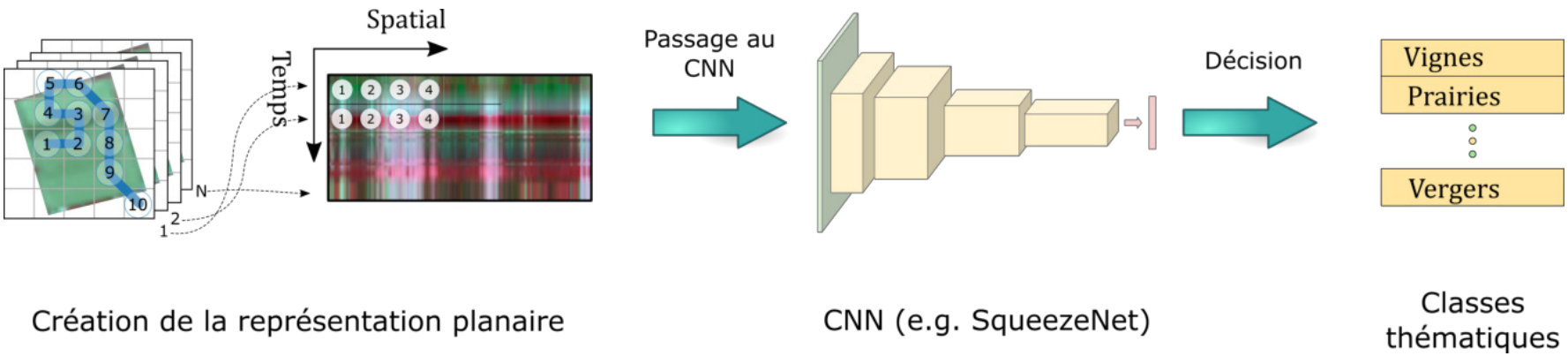
- Réduire la complexité de la structure des données : 2D+t à 2D
- **Transformation de chaque image en un vecteur 1D**

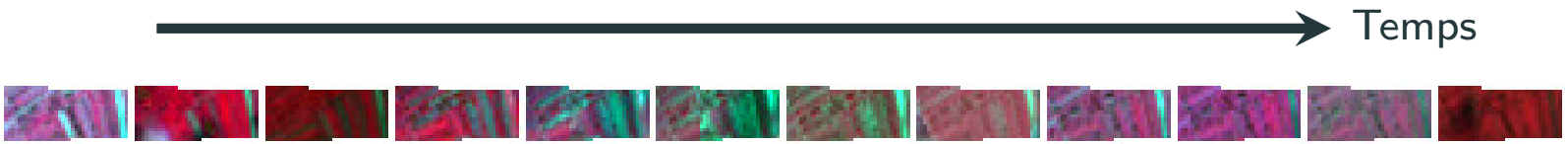
Problème : Perte partielle de l'information spatiale

⇒ Courbes globales ou courbes remplissant l'espace

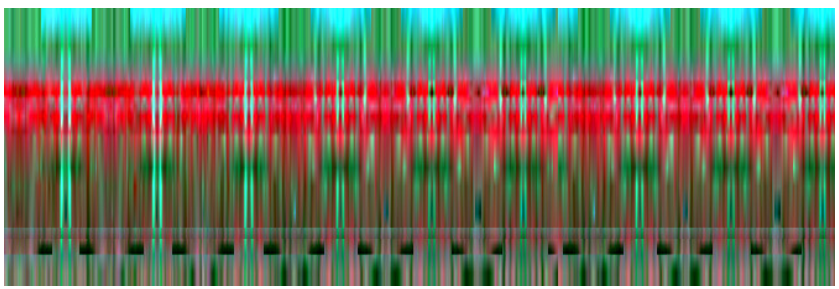
⇒ Courbes locales : *Random Walk (RW)*

Apprendre à étiqueter les STR

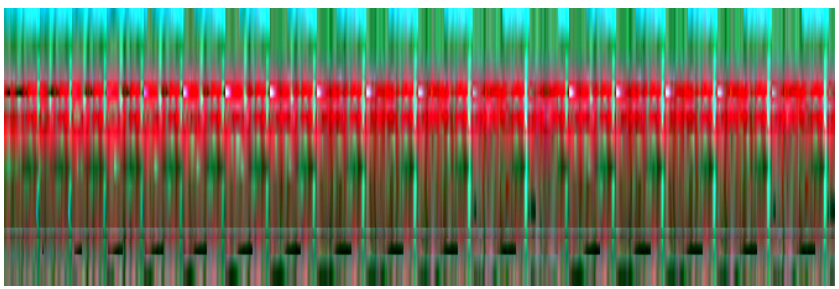




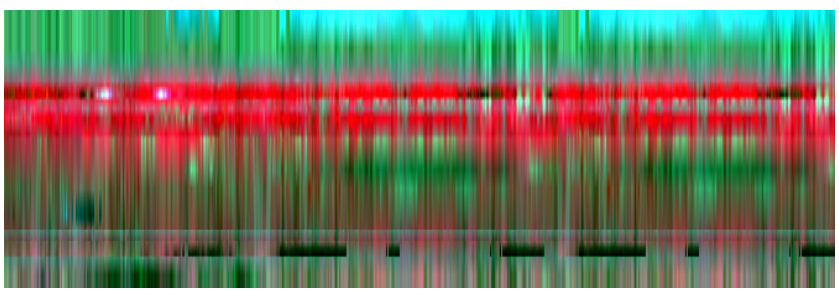
(a) Verger traditionnel



(b) $\Re_{snake}$

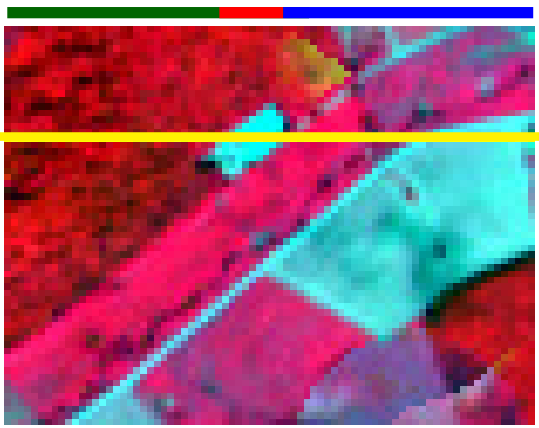


(c) $\Re_{spiral}$

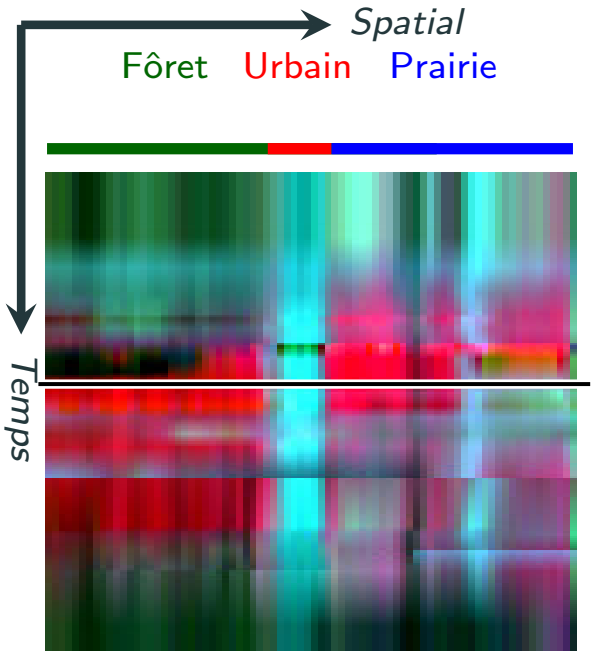


(d) $\Re_{Hilbert}$

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



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

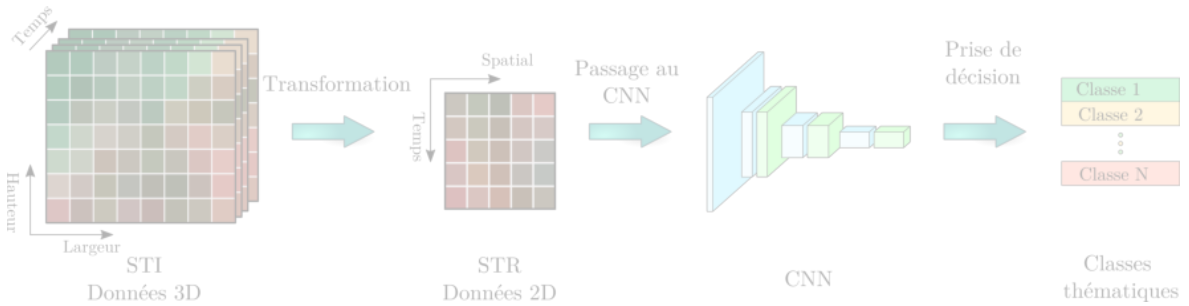


(c) STR associée au segment jaune dans (b)

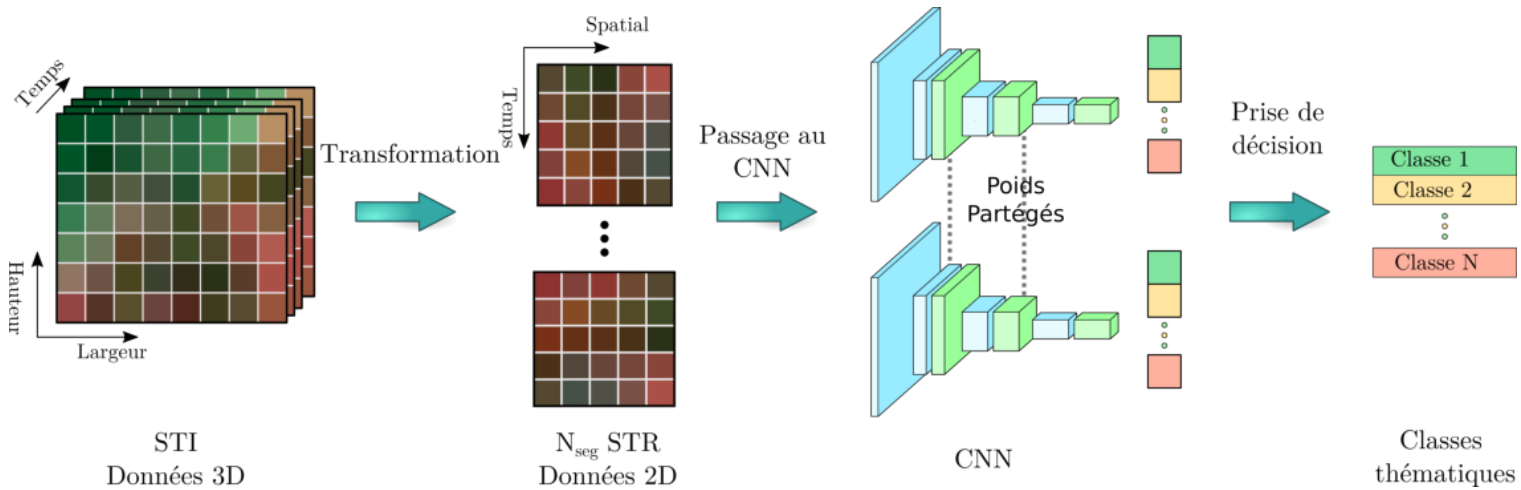
(d) Taille réelle de l'image Sentinel-2

DEEP-STAR : composée de deux étapes

- Hors-ligne



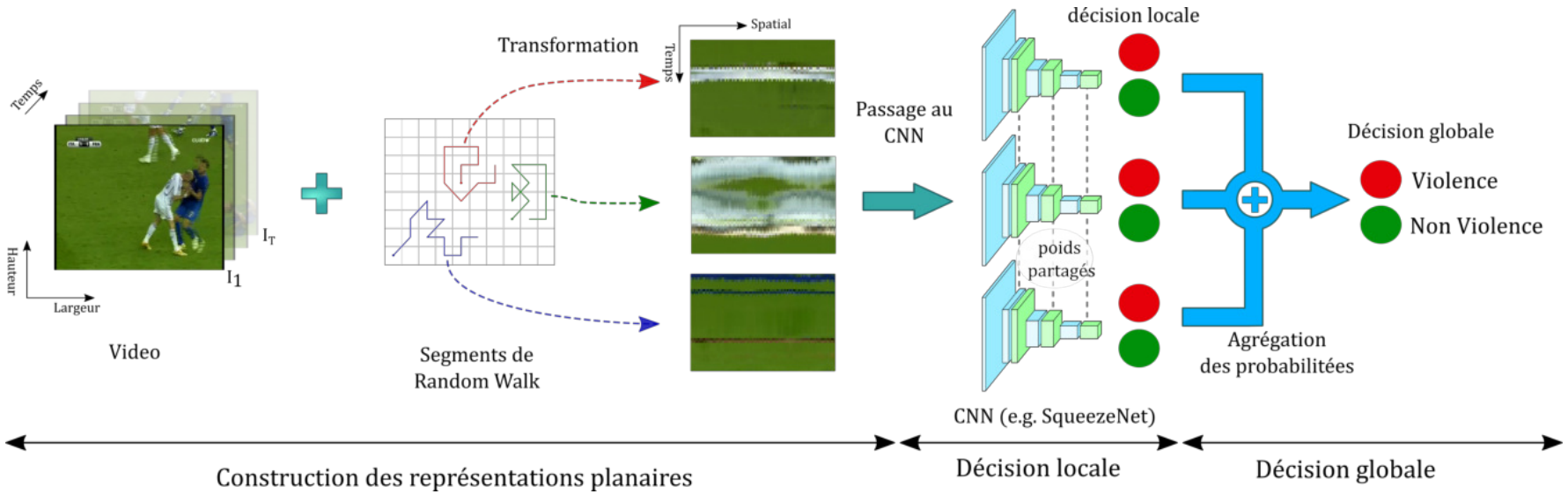
- En-ligne



Université de Paris

Étape d'inférence

- Génération de N_{seg} STR pour chaque STI
- Décision locale : classification de chaque STR
- Décision globale : agrégation des probabilités de toutes les STR



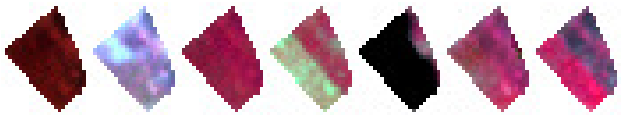
Application et évaluation sur deux cadres applicatifs :

- Application de télédétection
  Analyse de parcelles agricoles
- Application de vidéo
  Reconnaissance de scènes de violence

Données pour l'analyse des parcelles agricoles

(a) Prairie

(b) Vigne



(c) Verger intensif

(d) Verger traditionnel

- Série Temporelle d'Images Satellitaires issue de Sentinel-2 de 2017
- Parcelles extraites du RPG [6]
- Correction des délimitations par photo-interprétation
- Nombre de parcelles

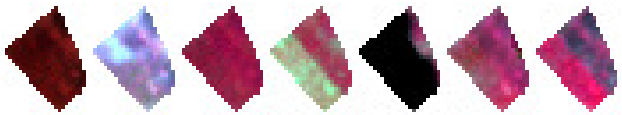
Classes	# poly.	aire (en pixels)	
		moyenne	écart-type
Prairies	1 045	250	338
Vignes	562	50	47
Vergers traditionnels	136	154	305
Vergers intensifs	191	129	115
Total	1 934	–	–

6. `http://professionnels.ign.fr/rpg`

Données pour l'analyse des parcelles agricoles

(a) Prairie

(b) Vigne



(c) Verger intensif

(d) Verger traditionnel

- Série Temporelle d'Images Satellitaires issue de Sentinel-2 de 2017
- Parcelles extraites du RPG [6]
- Correction des délimitations par photo-interprétation
- Nombre de parcelles

Classes	**# poly.**	**Méthode globale : G-STR**	**Méthode locale : MS-STR**			
		$\Re_*$	$RW_{10\%}$	$RW_{20\%}$	$RW_{50\%}$	$RW_{70\%}$
Prairies	1 045	1 757	26 110	51 688	128 424	179 914
Vignes	562	577	3 060	5 821	14 137	19 853
Vergers traditionnels	136	189	2 146	4 222	10 474	14 672
Vergers intensifs	191	226	2 564	5 027	12 414	17 408
Total	1 934	2 749	33 880	66 758	165 449	231 847

6. `http://professionnels.ign.fr/rpg`

Classification des parcelles à partir de STIS

- **STR crées avec la méthode globale (notée $G - STR$)**
- Modèle : SqueezeNet (Iandola et al., 2016)

Classification des parcelles à partir de STIS

- **STR crées avec la méthode globale (notée $G - STR$)**
- Modèle : SqueezeNet (Iandola et al., 2016)

TC : taux de bonne classification
ET : écart-type

Représentation		initialisation aléatoire TC	initialisation aléatoire ET	*Fine tuning* TC	*Fine tuning* ET
$\Re_{snake}$	sans AD	**70.00**	**1.70**	79.94	2.06
$\Re_{spiral}$		68.92	2.50	77.23	1.42
$\Re_{Hilbert}$		69.23	2.82	**81.69**	**1.88**

Classification des parcelles à partir de STIS

- **STR crées avec la méthode globale (notée $G - STR$)**
- Modèle : SqueezeNet (Iandola et al., 2016)

TC : taux de bonne classification
ET : écart-type

Représentation		initialisation aléatoire TC	initialisation aléatoire ET	*Fine tuning* TC	*Fine tuning* 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**

Classification des parcelles à partir de STIS

- **STR crées avec la méthode locale (notée $MS - STR$)**
- Modèle : SqueezeNet (Iandola et al., 2016)

TC : taux de bonne classification
ET : écart-type

	N_{seg}	Initialisation aléatoire		*Fine tuning*	
Représentation	**Entraîn. / Test**	**TC**	**ET**	**TC**	**ET**
$RW(10)$	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**

Université de Paris

Classification des parcelles à partir de STIS

- **STR crées avec la méthode locale (notée $MS - STR$)**
- Modèle : SqueezeNet (Iandola et al., 2016)

TC : taux de bonne classification
ET : écart-type

	N_{seg}	Initialisation aléatoire		*Fine tuning*	
Représentation	**Entraîn. / Test**	**TC**	**ET**	**TC**	**ET**
$RW(10)$	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**

Université de Paris

Classification des parcelles à partir de STIS

- **STR crées avec la méthode locale (notée $MS - STR$)**
- Modèle : SqueezeNet (Iandola et al., 2016)

TC : taux de bonne classification
ET : écart-type

Représentation	N_{seg} Entraîn. / Test	Initialisation aléatoire TC	Initialisation aléatoire ET	*Fine tuning* TC	*Fine tuning* ET
$RW(10)$	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**

Classification des parcelles à partir de STIS

- Comparaison avec les méthodes de l'état-de-l'art (EA)

TC : taux de bonne classification
ET : écart-type

Scores EA VS. nos meilleurs scores

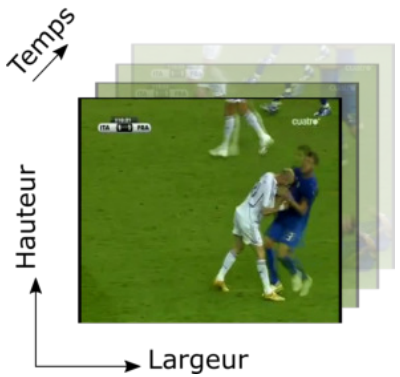
Méthodes	TC	ET
MS-STR $RW(50)_{70\%}$	**94.80**	**1.57**
TempCNN (Pelletier, Webb et Petitjean, 2019)	92.98	0.89
G-STR $\Re_{Hilbert}$	91.69	0.91
baML (Mauro et al., 2017)	91.25	0.53
3*D*-SqueezeNet (Köpüklü et al., 2019)	85.33	1.19
LSTM (Ienco et al., 2017)	83.48	2.29
ConvLSTM (Russwurm et Körner, 2018)	74.66	1.56

Application et évaluation sur deux cadres applicatifs :

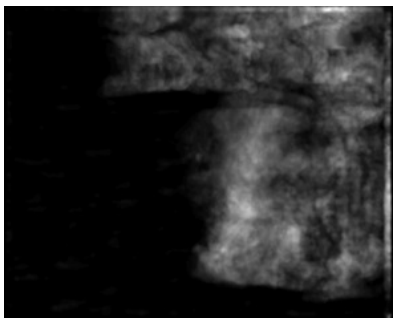
- Application de télédétection
  Analyse de parcelles agricoles
- Application de vidéo
  Reconnaissance de scènes de violence

Stratégie d'étiquetage des STR

- Les STR violentes sont générées dans les zones à grand flux optique



(a) Vidéo originale



(b) Flux moyen

(c) Binarisation du flux moyen

(d) ROI obtenue après post-traitement

- STR dans le domaine violent des vidéos violentes
- STR dans tous le domaine de l'image dans les vidéos non violentes + STR en-dehors du domaine violent dans les vidéos violentes

	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

Classification des vidéos pour la reconnaissance des scènes de violence

- Modèle : SqueezeNet (Iandola et al., 2016)
- Méthode utilisée : MS-STR – RW(100)

Université de Paris

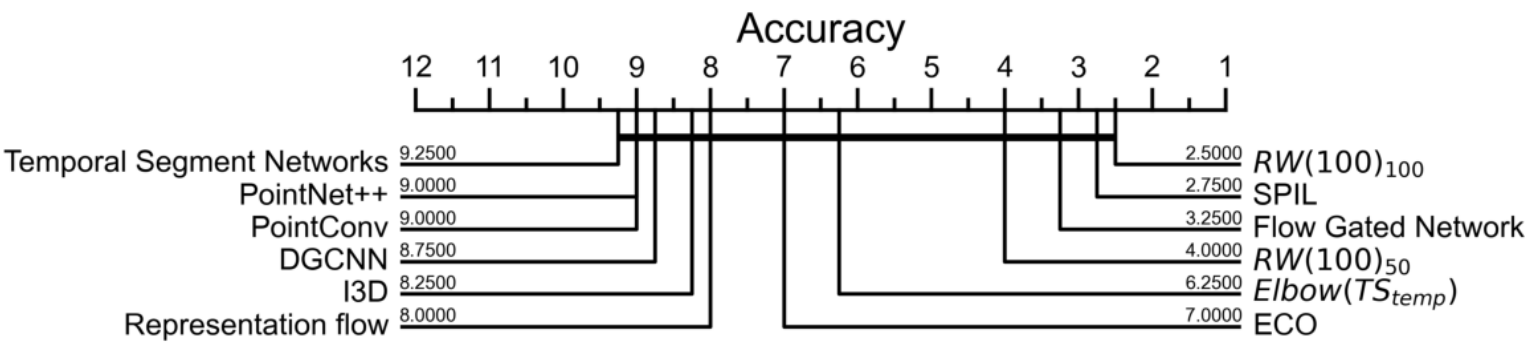
Classification des vidéos pour la reconnaissance des scènes de violence

- Modèle : SqueezeNet (Iandola et al., 2016)
- Méthode utilisée : MS-STR – RW(100)

Méthode	*RWF2000*	*Movies fights*	*Hockey fights*	*Crowd Violence*
Deep-STaR				
$N_{seg}^{test} = 100$	**93.8**	**98.5**	94.4	89.8
CNN 3D				
Temp. Seg. Nets (Wang et al., 2016)	81.5	94.2	91.5	81.5
I3D (Carreira et Zisserman, 2017)	83.4	95.8	93.4	83.4
Represent. flow (Wang et al., 2017)	85.3	97.3	92.5	85.9
Flow Gated Net (Cheng, Cai et Li, 2020)	87.3	n/a	**98.0**	88.8
ECO (Zolfaghari, Singh et Brox, 2018)	83.7	96.3	94.0	84.7
Nuage de points				
PointNet++ (Qi et al., 2017)	78.2	89.2	89.7	89.2
PointConv (Wu, Qi et Li, 2019)	76.8	91.3	89.7	89.2
DGCNN (Wang et al., 2019)	80.6	92.6	90.2	87.4
SPIL (Su et al., 2020)	89.3	**98.5**	96.8	**94.5**

Classification des vidéos pour la reconnaissance des scènes de violence

- Modèle : SqueezeNet (Iandola et al., 2016)
- Méthode utilisée : MS-STR – RW(100)

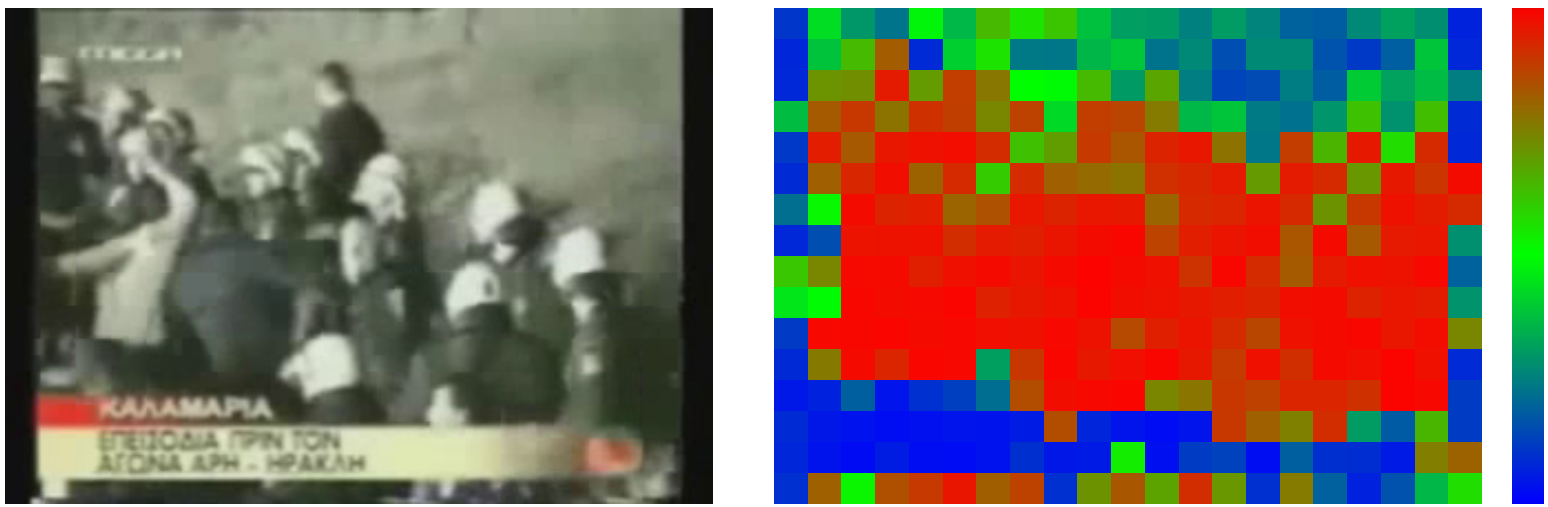


Méthode	*RWF2000*	*Movies fights*	*Hockey fights*	*Crowd Violence*
Flow Gated Net (Cheng, Cai et Li, 2020)	87.3	n/a	**98.0**	88.8
ECO (Zolfaghari, Singh et Brox, 2018)	83.7	96.3	94.0	84.7
Nuage de points				
PointNet++ (Qi et al., 2017)	78.2	89.2	89.7	89.2
PointConv (Wu, Qi et Li, 2019)	76.8	91.3	89.7	89.2
DGCNN (Wang et al., 2019)	80.6	92.6	90.2	87.4
SPIL (Su et al., 2020)	89.3	**98.5**	96.8	**94.5**

Reconnaissance de scènes de violence

Réalisation d'une carte de probabilités pour localiser la violence

- Diviser le domaine spatial de la vidéo selon une grille
- Générer une STR par cellule de la grille
- Indiquer la probabilité de violence de chaque STR dans chaque cellule



(a) **(b)**

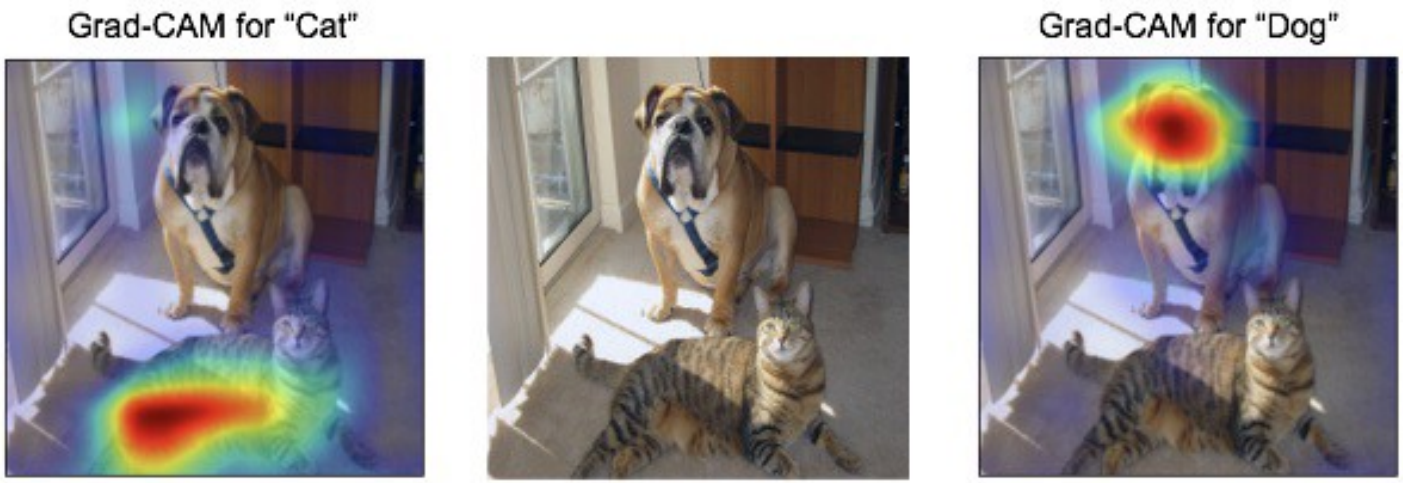
Résultat de la classification d'une vidéo issue de la base Crowd Violence : (a) Une image de la vidéo ; (b) Carte des probabilités de violence (échelle de couleur : rouge (forte violence), bleu (pas de violence)).

Étude des variations des séquences temporelles d'images

Explicabilité des décisions du CNN

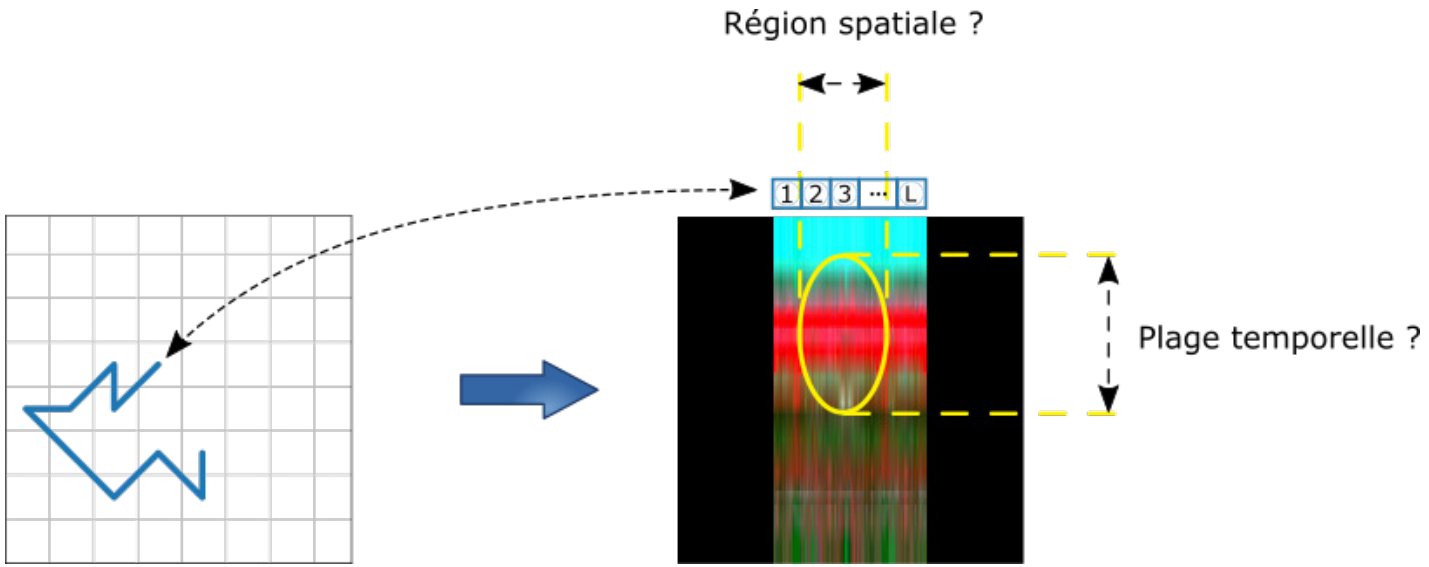
Interprétation visuelle

- Explication de la décision prise par le CNN
- 2 stratégies co-existent : *trainable* attention vs. **post-hoc attention**
- Visualisation des régions qui caractérisent chaque classe
- **Solution :** utilisation du Grad-CAM++ (Chattopadhyay et al., 2018) qui est une amélioration des *Class Activation Map*



Interprétation visuelle

- Explication de la décision prise par le CNN
- 2 stratégies co-existent : *trainable* attention vs. **post-hoc attention**
- Visualisation des régions qui caractérisent chaque classe
- **Solution :** utilisation du Grad-CAM++ (Chattopadhyay et al., 2018) qui est une amélioration des *Class Activation Map*



- Attention temporelle : choisir une plage temporelle significative
- Attention spatiale : explicabilité de la décision dans l'espace spatial original

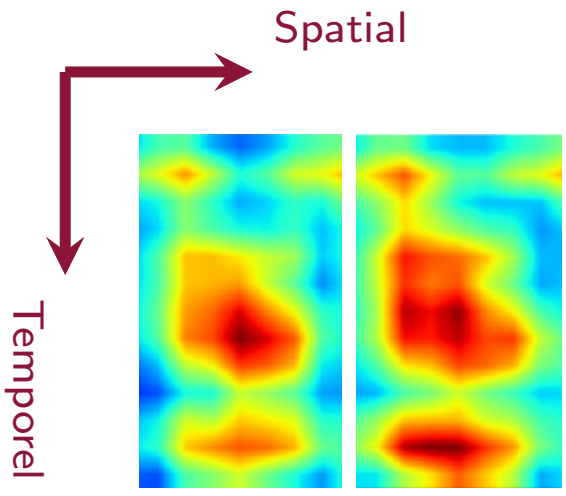
Attention temporelle

Objectif

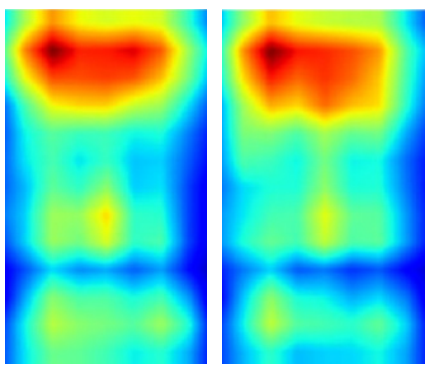
- Identification du domaine temporel le plus discriminant de chaque classe

Questions

- Est-il essentiel d'analyser tout le domaine temporel ?
- Peut-on améliorer les scores en n'utilisant qu'une partie du domaine temporel ?



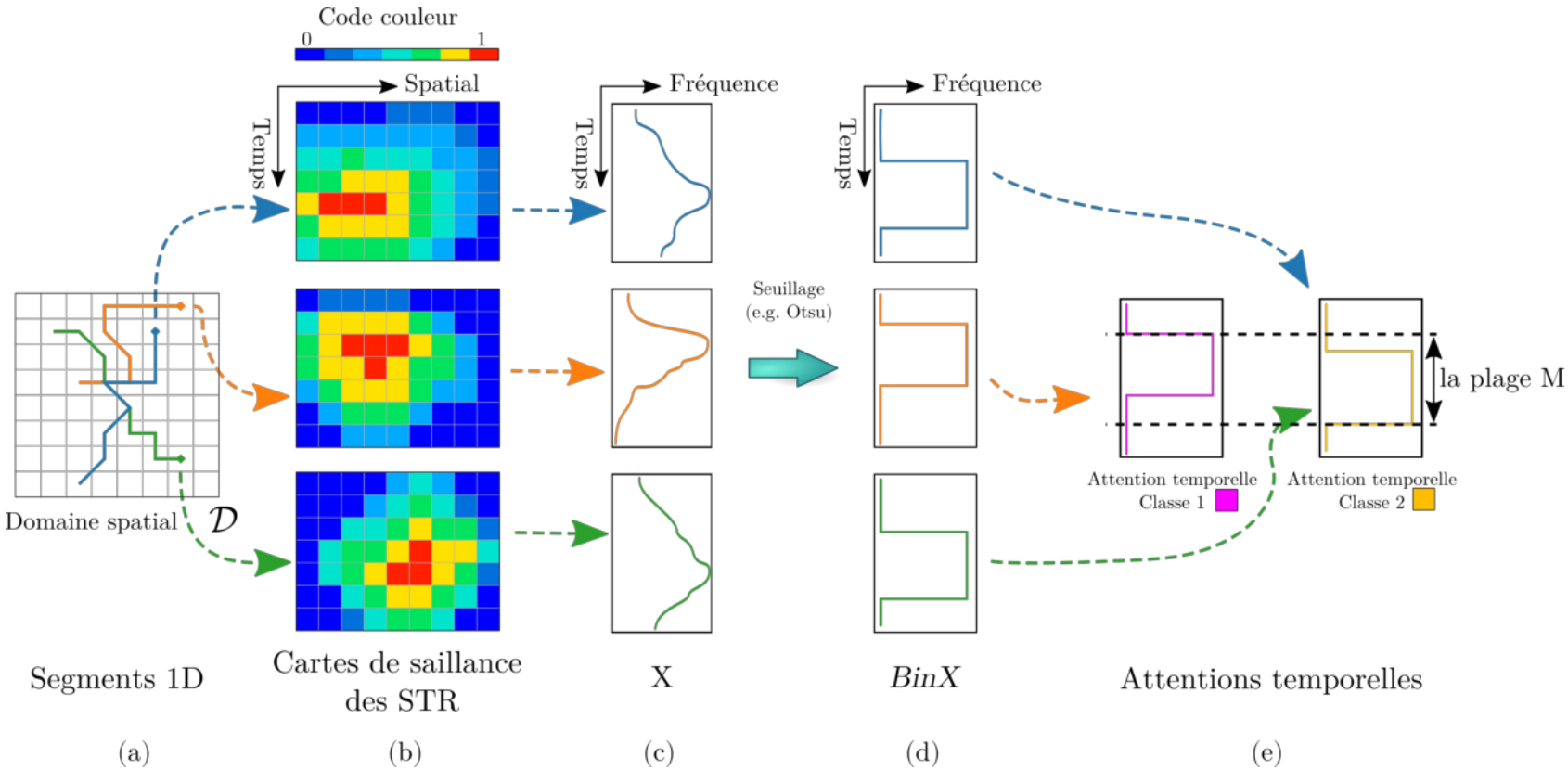
(a) $S^{c(STR)}$ de Prairie



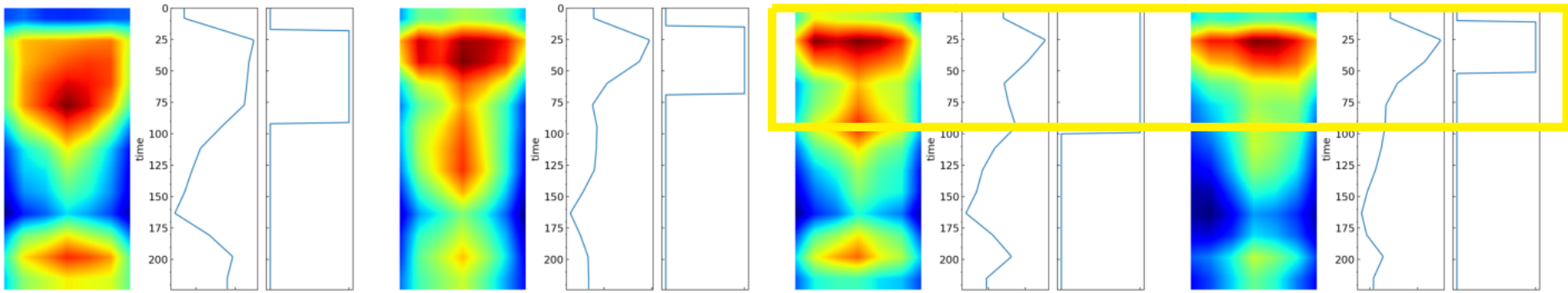
(b) $S^{c(STR)}$ de Verger intensif

Attention temporelle

- Capturer l'attention temporelle X
- Binariser le profil de l'attention temporelle, noté $BinX$
- Définir un masque pour capturer le domaine temporel le plus significatif



Attention temporelle avec DEEP-STAR



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

- DEEP-STAR : nouvelle plage temporelle [0 ; 120]

	224 dates		Nouvelle plage temporelle	
	TC	**ET**	**TC**	**ET**
RW(100)	93.00	2.44	94.00	2.54

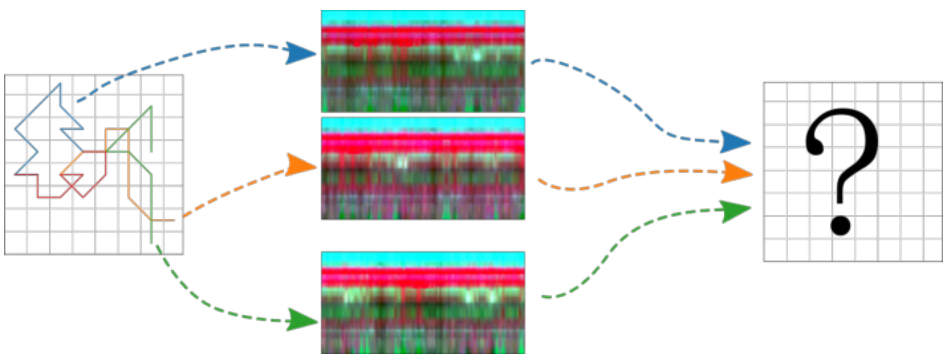
Attention spatiale

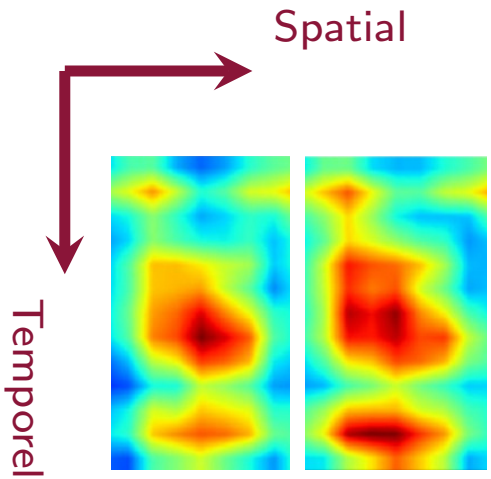
Objectif

- Explication de la décision en mettant en évidence les décisions spatiales intéressantes dans l'espace image original 2D de la STI

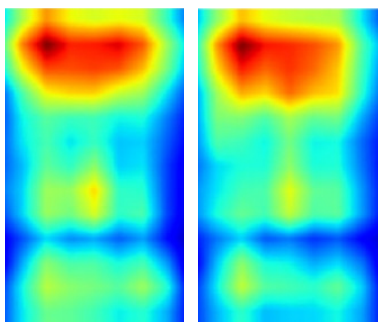
Questions

- Comment revenir vers l'espace original en considérant les N_{seg} représentations ?





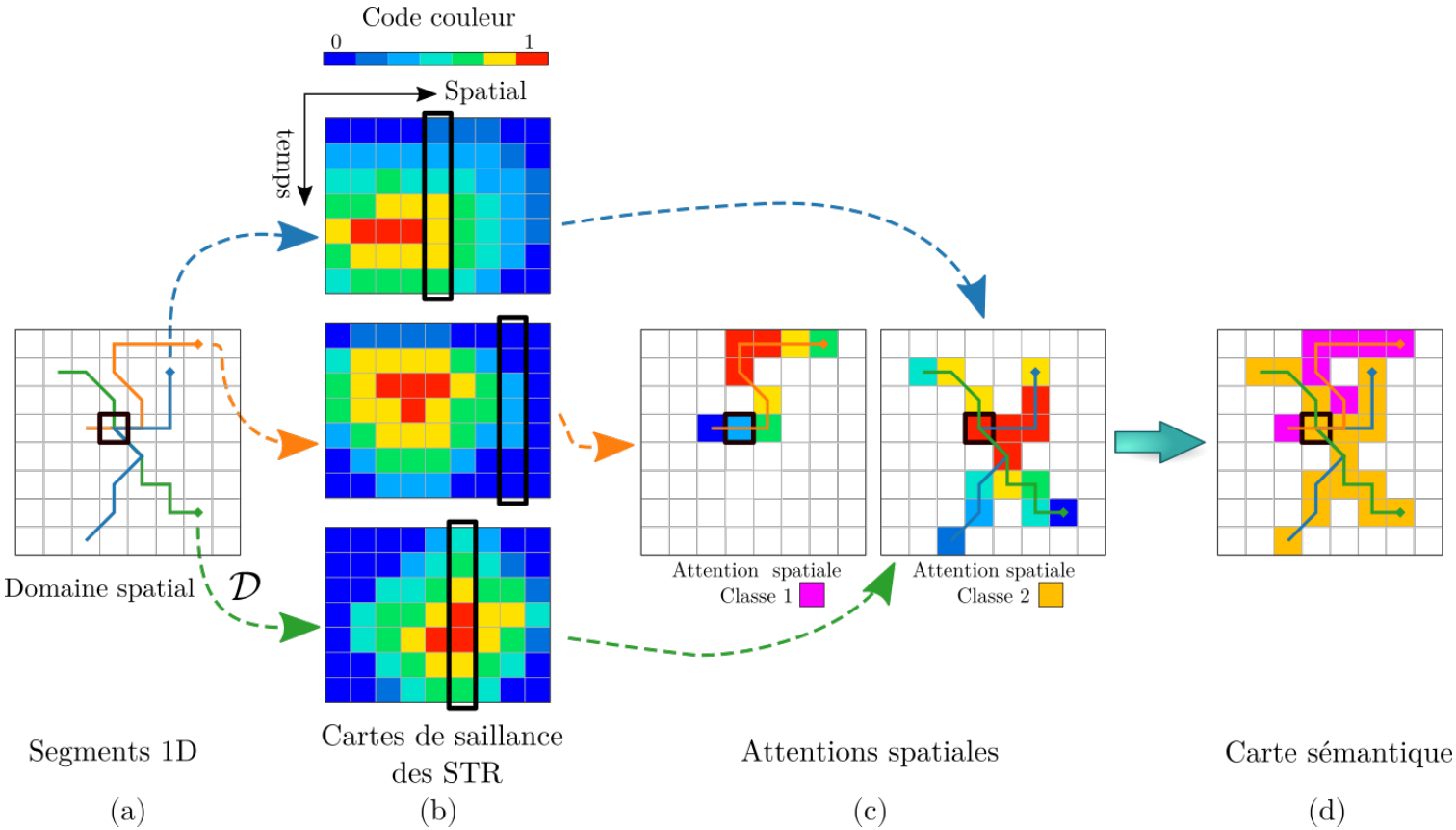
(a) $S^{c(STR)}$ de Prairie



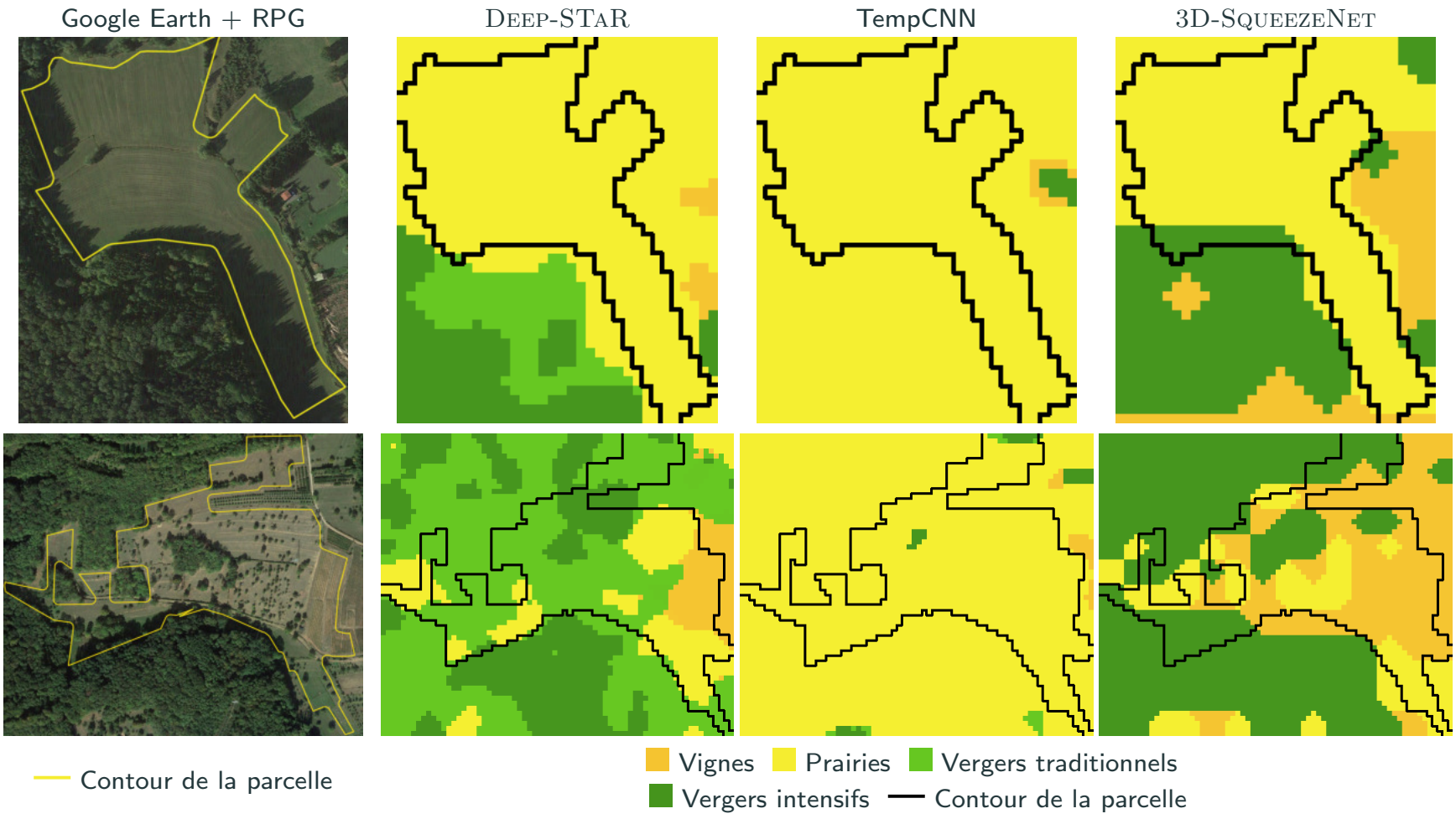
(b) $S^{c(STR)}$ de Verger intensif

Attention spatiale

- Capturer l'attention spatiale d'un pixel p
- Attribuer l'attention maximale du pixel p entre toutes les STR à la position spatiale
- Créer une carte de segmentation sémantique en affectant une couleur à chacune des attentions maximales des C classes



Attention spatiale



Étude des variations des séquences temporelles d'images

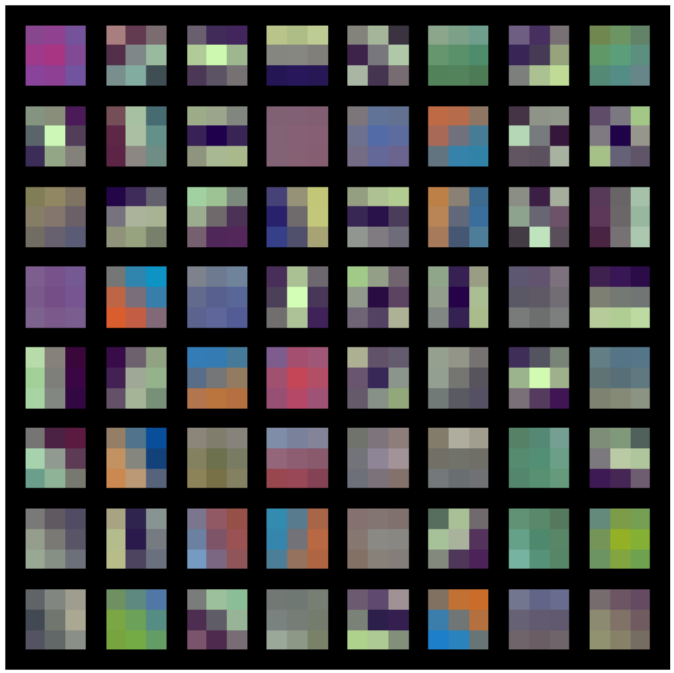
Analyse des filtres du CNN

Analyse des filtres du CNN

- En fonction des informations portées par les STR

  $\Rightarrow$ Quelles sont les informations les plus utilisées par le CNN ?

- **Proposition** : alimenter le CNN avec des images synthétiques

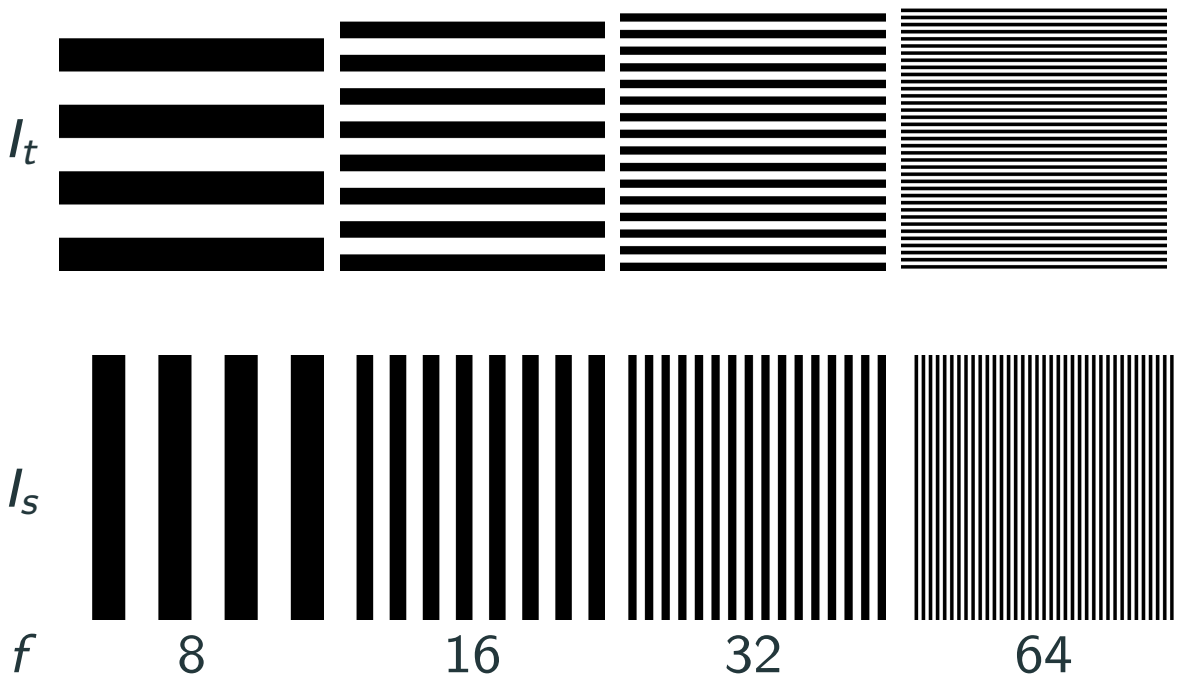


Analyse des filtres du CNN

- En fonction des informations portées par les STR

  $\Rightarrow$ Quelles sont les informations les plus utilisées par le CNN ?

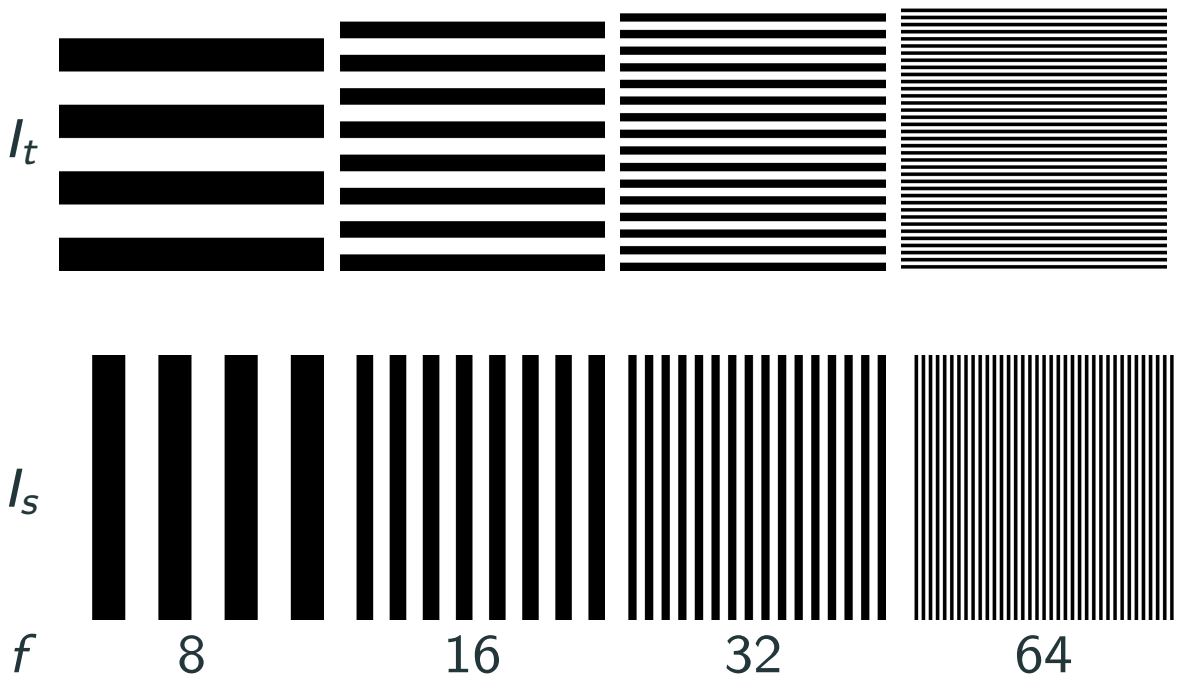
- **Proposition** : alimenter le CNN avec des images synthétiques



Analyse des filtres du CNN

- Calculer l'énergie (E) des k réponses de la couche considérée
- Calculer le ratio des k énergies

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

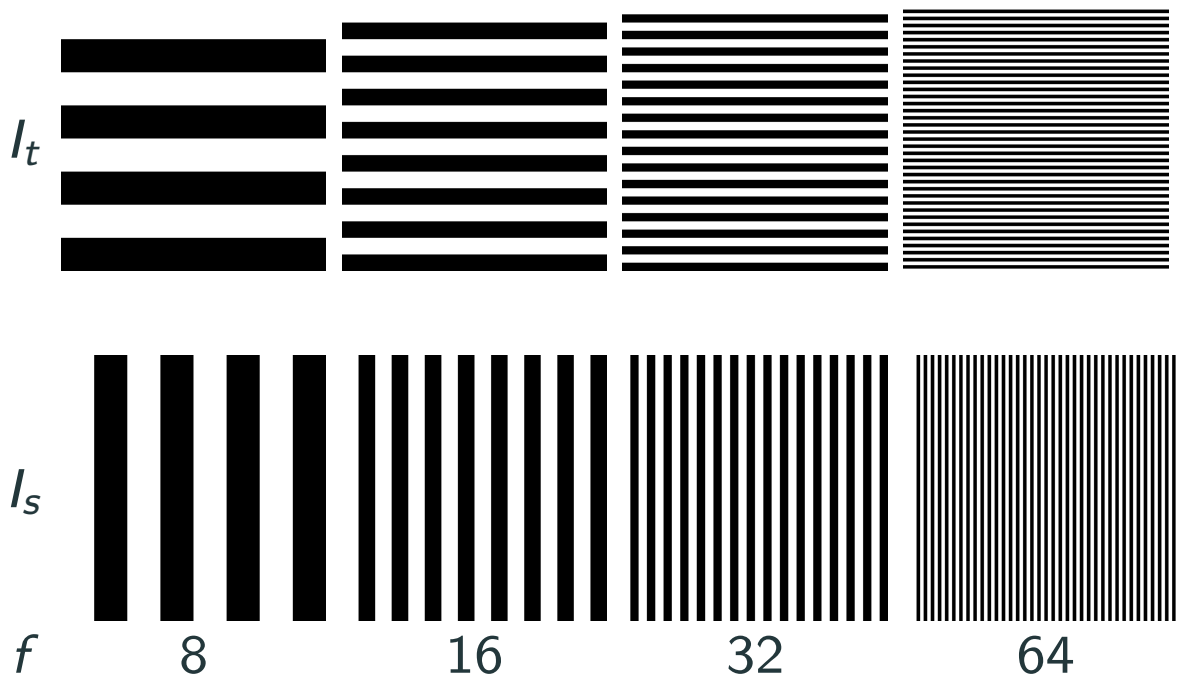


Université de Paris

Analyse des filtres du CNN

Interprétation des ratios R_{st}

- **Les filtres spatiaux** : le rapport $R_{st}(k)$ est supérieur à $1 + \mu$
- **Les filtres temporels** : 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 $1 + \mu$

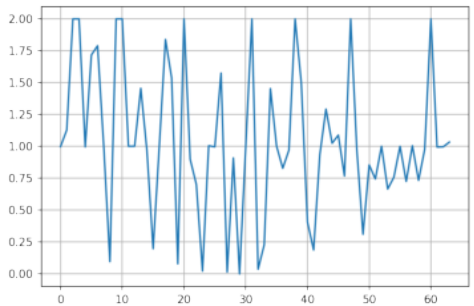


Analyse des filtres de SqueezNet entraîné sur les STR de parcelles

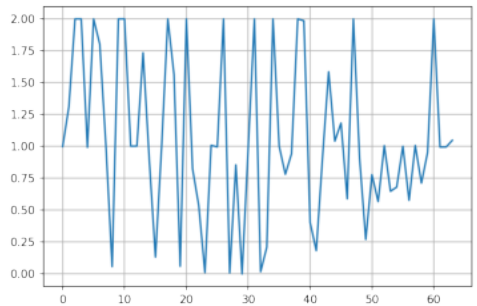
- Meilleur modèle : **MS-STR** – $RW(50)_{70\%}$

Analyse des filtres de SqueezNet entraîné sur les STR de parcelles

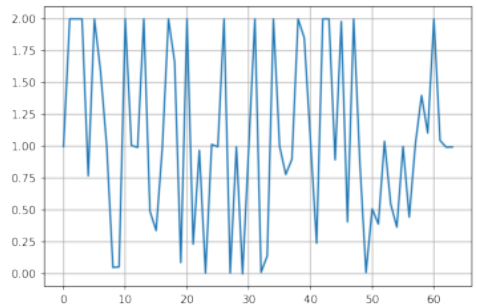
- Meilleur modèle : **MS-STR** – $RW(50)_{70\%}$
- Visualisation des k ratios des énergies



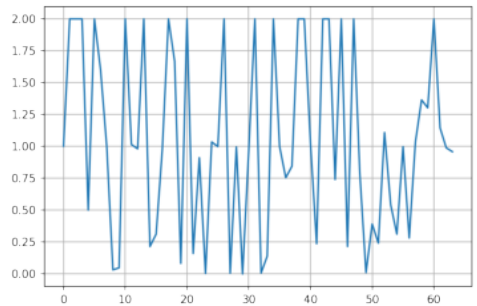
$f = 8$



$f = 16$



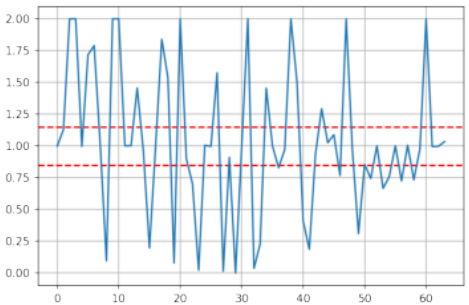
$f = 32$

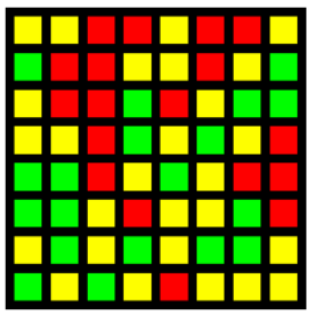


$f = 64$

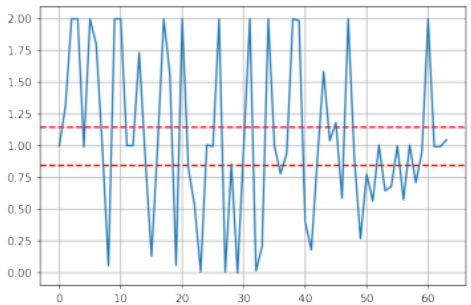
Analyse des filtres de SqueezNet entraîné sur les STR de parcelles

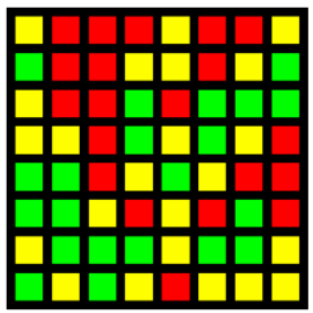
- Meilleur modèle : **MS-STR** – $RW(50)_{70\%}$
- Visualisation des k ratios des énergies
- Classification des filtres, $\mu = \nu = 0.15$



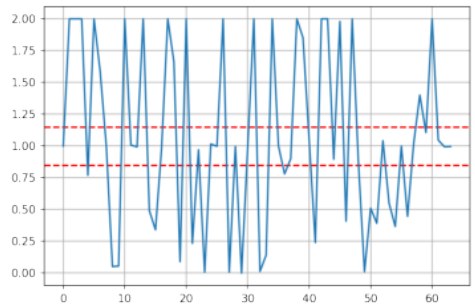


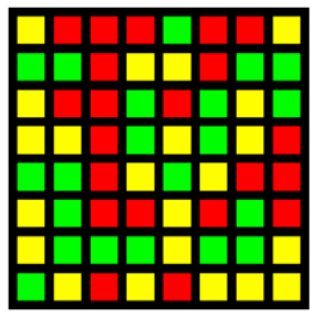
$f = 8$



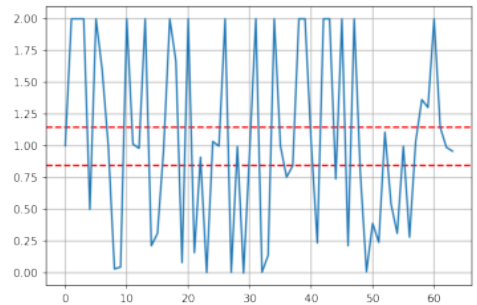


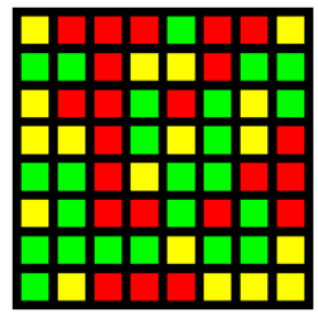
$f = 16$





$f = 32$





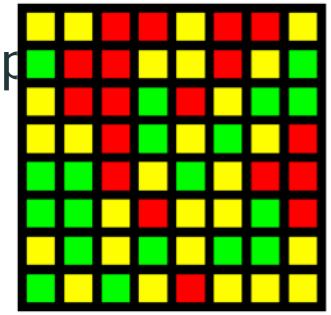
$f = 64$

Filtre temporel Filtre spatial Filtre spatio-temporel

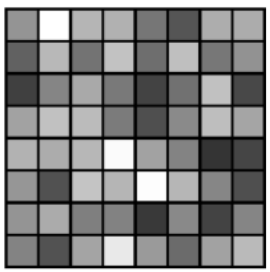
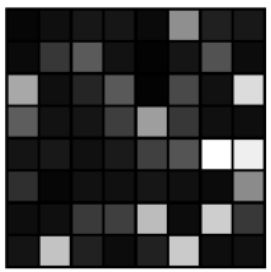
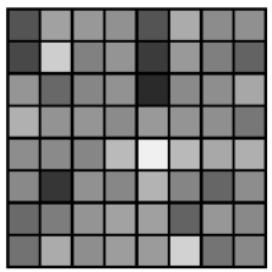
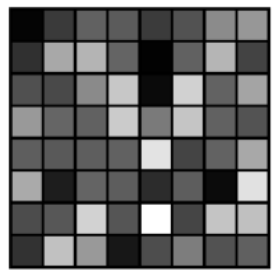
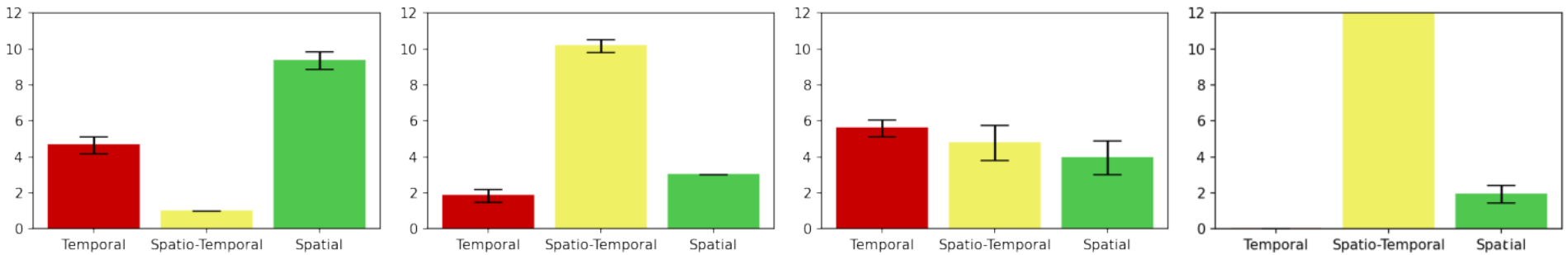
Université de Paris

Analyse des filtres de SqueezNet entraîné sur les STR de p

- Application sur le meilleur modèle : **MS-STR** – $RW(50)_{70\%}$
- Classification des filtres, $\mu = \nu = 0.15$
- Analyse des filtres pour quelques parcelles



$f = 8$

(c) Verger intensif **(d)** Verger traditionnel **(e)** Prairie **(f)** Vigne

Ce qu'il faut retenir

- Proposition d'une représentation spatio-temporelle
  - Utilisation d'un réseau de neurones convolutif 2D pour apprendre des caractéristiques spatio-temporelles
  - Utilisation d'un modèle pré-entraîné (e.g. la base ImageNet)
- Proposition d'un mécanisme d'attention *post-hoc*
  - Pour analyser le domaine temporel
  - Pour générer une carte de segmentation sémantique
- Analyse de la nature de l'information portée par les filtres

Conclusion et perspectives

Conclusion

- Proposition de 2 méthodes pour la définition de caractéristiques
  - Caractéristique artisanale qui mesure la stabilité temporelle
  - Caractéristiques apprises automatiquement grâce à un réseau de neurones convolutif 2D
- Utilisation de chacune des méthodes dans 2 cadres applicatifs
  - Analyse de la couverture urbaine
  - Analyse de parcelles agricoles
  - Analyse des vidéos pour la reconnaissance des scènes de violences

Perspectives

Mesure de la stabilité

- Étudier l'égalité vectorielle (e.g. RGB)
- Étudier la quantification des valeurs
- Étudier les différentes relaxations
- Utiliser les méthodes d'apprentissage pour :
  - Générer l'image du résumé
  - Classifier la séquence avec un modèle à 2 têtes (une pour le résumé et l'autre pour la séquence)

Perspectives

Étude des variations des séquences temporelles d'images

- Utiliser un autre modèle que SqueezeNet
- Pour l'analyse de vidéos
  - Étudier la zone de violence
  - Ajouter une troisième classe d'arrière plan
  - Classifier l'image des probabilités pour avoir une décision globale
- Appliquer les mécanismes d'attention pour l'analyse de vidéos

Appliquer les deux méthodes proposées dans d'autres domaines tels que le biomédical

Merci pour votre attention

Code source : `https://github.com/mchelali/TemporalStability`

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

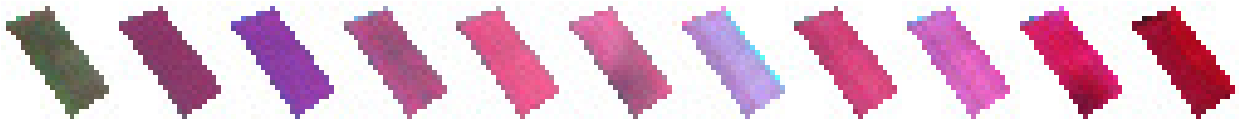
Mohamed CHELALI

26 novembre 2021

Préparation des STR de parcelles

- Modèle : SqueezeNet $\Rightarrow$ Taille d'entrée : 224×224

Adaptation des STR à la taille d'entrée

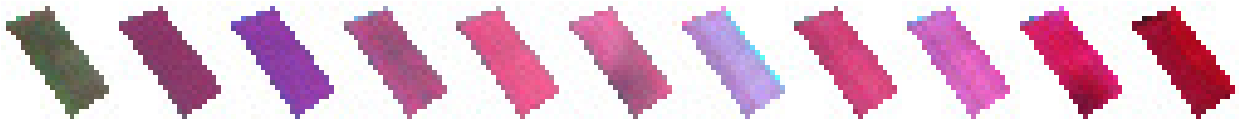


- méthode G-STR

Préparation des STR de parcelles

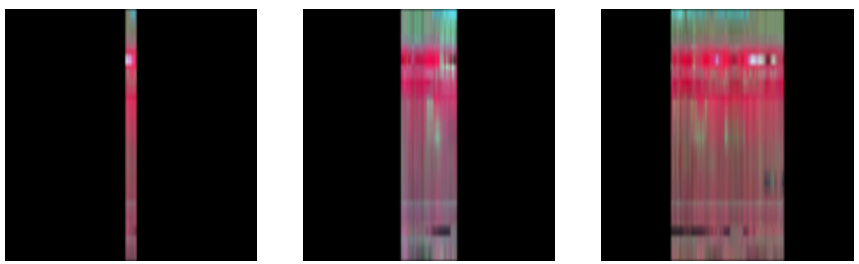
- Modèle : SqueezeNet ⇒ Taille d'entrée : 224×224

Adaptation des STR à la taille d'entrée



- méthode G-STR

- méthode MS-STR



(a) RW(10) **(b)** RW(50) **(c)** RW(100)

Classification des parcelles à partir de STIS

- Comparaison avec les méthodes de l'état-de-l'art (EA)

Temps d'inférence

Méthodes	Temps moyen (en secondes)
G-STR $\Re_{Hilbert}$	2.72
baML (Mauro et al., 2017)	11.91
TempCNN (Pelletier, Webb et Petitjean, 2019)	13.50
MS-STR $RW(50)$	22.57
ConvLSTM (Russwurm et Körner, 2018)	23.16
3D-SqueezeNet (Köpüklü et al., 2019)	26.50
LSTM (Ienco et al., 2017)	26.96

Classification des vidéos pour la reconnaissance des scènes de violence

- Modèle : SqueezeNet (Iandola et al., 2016)
- Méthode utilisée : MS-STR – RW(100)

Méthode	*RWF2000*	*Movies fights*	*Hockey fights*	*Crowd Violence*
Deep-STaR				
$N_{seg}^{test} = 100$	**93.8±0.52**	98.5±1.27	94.4±3.93	89.8±2.38
CNN 3D				
Temp. Seg. Nets (Wang et al., 2016)	81.5	94.2	91.5	81.5
I3D (Carreira et Zisserman, 2017)	83.4	95.8	93.4	83.4
Represent. flow (Wang et al., 2017)	85.3	97.3	92.5	85.9
Flow Gated Net (Cheng, Cai et Li, 2020)	87.3	n/a	**98.0**	88.8
ECO (Zolfaghari, Singh et Brox, 2018)	83.7	96.3	94.0	84.7
Nuage de points				
PointNet++ (Qi et al., 2017)	78.2	89.2	89.7	89.2
PointConv (Wu, Qi et Li, 2019)	76.8	91.3	89.7	89.2
DGCNN (Wang et al., 2019)	80.6	92.6	90.2	87.4
SPIL (Su et al., 2020)	89.3	**98.5**	96.8	**94.5**