Réseau Convolutif Spatio-Temporel 3D pour la Reconnaissance Précoce de Gestes Manuscrits Non-Segmentés

William Mocaër, Eric Anquetil, Richard Kulpa

Réseau Convolutif Spatio-Temporel 3D pour la Reconnaissance Précoce de Gestes Manuscrits Non-Segmentés

W. Mocaër[1] E. Anquetil[1] R. Kulpa[2]

[1] Univ Rennes, CNRS, IRISA
[2] Univ Rennes, Inria, M2S

william.mocaer@irisa.fr

Résumé

La reconnaissance précoce de gestes manuscrits non segmentés est la tâche consistant à reconnaître le plus rapidement possible des gestes dessinés dans un flux continu, les uns après les autres. Cette tâche est particulièrement difficile pour les gestes multi-touch car il n'est pas possible de savoir quand le geste a commencé et s'est terminé. Pour les gestes mono-stroke, dans un contexte d'application où le doigt n'est jamais retiré de l'appareil entre les gestes, la reconnaissance est encore plus complexe. Dans ce travail, nous présentons une extension du réseau convolutif 3D à long terme (OLT-C3D) afin d'aborder la nouvelle tâche de reconnaissance précoce des gestes non segmentés qui n'a été abordée que par très peu de travaux jusqu'à présent. Pour évaluer notre approche, nous avons créé deux bases de données synthétiques à partir de bases existantes librement disponibles, MTGSetB et ILGDB, en simulant les données en continu dans deux scénarios d'application différents. Nous proposons également une nouvelle métrique pour évaluer cette tâche spécifique. Notre approche obtient de bonnes performances sur les deux nouveaux jeux de données et servira de référence pour de futurs travaux sur cette tâche.

Mots Clef

Réseau à Convolution Spatio-temporel, Reconnaissance précoce, gestes non segmentés, flux de données, Online Long-Term C3D, WaveNet 3D

Abstract

Early recognition of untrimmed handwritten gestures is the task of recognizing as soon as possible gestures drawn in a continuous stream, the ones after the others. This is particularly challenging for multi-touch gestures because it is not possible to know when the gesture has started and finished. For mono-stroke gestures, in an application context where the finger is never removed from the device between gestures, the recognition is even more complex. In this work we present an extension of the Online Long-Term Convolutional 3D (OLT-C3D) network to address the new task of early recognition of untrimmed gestures which have been addressed by very few works. To evaluate our approach, we created two synthetic datasets from freely available benchmarks, MTGSetB and ILGDB, simulating the streaming data in two different application scenarios. We also propose a new metric to evaluate this specific task. Our approach achieves good performances on the two new datasets and will be a baseline for future works on this challenging task.

Keywords

Spatio-Temporal Convolutional Neural Network Early recognition Untrimmed gestures Streaming data Online Long-Term C3D WaveNet 3D.

1 Introduction

Du point de vue de l'interaction avec l'utilisateur, des interactions réactives et naturelles avec les dispositifs tactiles sont essentielles pour une expérience réussie. L'interaction gestuelle permet à l'utilisateur de manipuler naturellement le dispositif, mais elle est souvent limitée à des fonctionnalités très basiques comme le zoom, la rotation et le défilement. La difficulté d'ajouter de nouveaux gestes est double : la précision de la reconnaissance et la réactivité du système. Avec plus de gestes, certains partageront des débuts communs, de sorte que le système ne peut pas toujours prédire à partir des premières traces, car cela conduirait à des exécutions de commandes potentiellement indésirables. Mais lorsque nous voulons manipuler le dispositif en temps réel, nous ne pouvons pas attendre que l'utilisateur finisse de faire le geste de zoom pour appliquer effectivement le zoom, le geste doit être détecté dès les premiers instants pour produire un retour en direct. Pour être en mesure de gérer de telles manipulations directes, nous avons besoin d'un système capable de reconnaître très tôt les gestes de l'utilisateur, juste après que la partie commune entre les gestes soit passée. Aujourd'hui, cela ne fonctionne que pour un nombre très limité de gestes bien conçus, avec une approche ad hoc qui ne peut être généralisée.
Une autre difficulté dans l'interaction de l'utilisateur avec le geste est un contexte d'application où les gestes sont faits l'un après l'autre, alors les limites de début et de fin

des gestes ne sont pas nécessairement claires et cela peut perturber la reconnaissance. Cela est particulièrement vrai pour les gestes mono-touch, lorsque le doigt n'est jamais retiré du dispositif entre les gestes, comme pour l'écriture d'un mot à la main, mais les gestes sont complètement mélangés dans l'espace (voir figure 1). Les gestes multi-stroke posent également problème car les pauses intra-gestes (lorsque tous les doigts sont retirés de l'appareil) entre deux tracés ne peuvent pas être distinguées d'une pause inter-gestes, donc nous ne sommes donc pas en mesure de dire où commence le geste. Peu de travaux s'intéressent à ce problème difficile, la plupart d'entre eux traitant de la reconnaissance précoce de gestes effectués par un corps humain complet, soit à partir de vidéos RVB [1] ou de vidéos de squelettes 3D [2–4]. En ce qui concerne la reconnaissance de gestes 2D, certains travaux ont abordé la reconnaissance précoce de gestes segmentés [5–7], mais pas dans un contexte non segmenté.

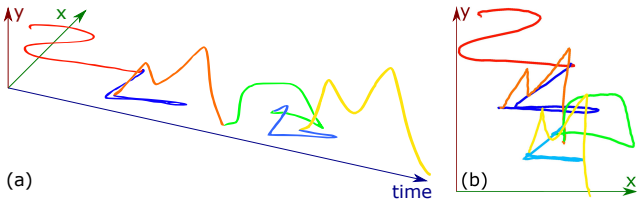


FIGURE 1 – Exemple d'une séquence non segmentée générée à partir d'ILGDB. Chaque couleur représente des gestes différents. Dans un contexte non segmenté, les gestes sont enchaînés dans le même espace. (a) : selon les trois dimensions, x, y et le temps. (b) : uniquement les dimensions spatiales.

Pour résoudre ce problème, nous présentons une approche basée sur un CNN 3D spatio-temporel appelé OLT-C3D [7] conçu pour la reconnaissance de gestes segmentés. OLT-C3D, couplé à un système de rejet temporel, est capable de reporter une détection pour éviter une détection mal classifiée, mais il peut être instable car la cohérence entre les images n'est pas explicitement entraînée, ce qui le rend peu fiable dans un contexte non segmenté. Nous l'étendons au contexte non segmenté en régularisant le réseau avec la perte CTC (Connectionist Temporal Classification), ce qui améliore la stabilité et la robustesse du système. Comme il s'agit d'une nouvelle tâche, nous définissons un nouveau protocole d'évaluation : nous avons créé deux ensembles de données artificielles pour simuler le flux de gestes, et nous proposons une nouvelle métrique d'évaluation.

Les contributions de cet article sont résumées comme suit :

— Nous avons conçu une méthode pour la tâche de reconnaissance précoce de gestes non segmenté en utilisant un CNN 3D spatio-temporel avec un système de rejet temporel. Pour améliorer sa stabilité de prédiction dans le temps, nous proposons de régulariser le réseau avec la perte CTC.

— Nous proposons deux stratégies de représentation des gestes indépendantes de la vitesse pour les gestes multi-touch et mono-stroke.

— Nous construisons deux nouveaux ensembles de données stimulantes de séquences de gestes générées pour évaluer notre système sur des scénarios applicatifs et nous proposons une nouvelle métrique pour évaluer spécifiquement la reconnaissance précoce des gestes non segmentés.

2 État de l'art

Peu de travaux ont abordé la tâche de reconnaissance précoce des **gestes 2D**. Uchida et al. [5] ont construit un système basé sur des classificateurs de trames multiples, un classificateur faible est construit par trame. À chaque instant, le système combine les résultats des classifieurs des instants précédents et de l'instant courant pour déterminer la classification courante. Une autre approche de combinaison de classificateurs est également utilisée dans le travail de Chen et al. [6], ils ont conçu des classifieurs dépendant de la longueur des gestes et un système de rejet basé sur les scores de confiance et la répétition de la prédiction. Yamagata et al. [8] ont conçu une approche modélisant explicitement les bifurcations de trajectoire entre les chiffres manuscrits, un réseau LSTM est utilisé pour prédire la classe et la trajectoire future. Récemment, une nouvelle approche basée sur un réseau de neurones convolutif 3D (CNN) appelé OLT-C3D (pour Online Long-Term Convolutional 3D) [7] a été conçue pour gérer la visibilité à long terme sans avoir besoin d'une couche de récurrence grâce à des convolutions dilatées dans le temps. Cette approche dispose d'un système de rejet pour éviter les erreurs de classification dans les premiers stades. Toutes ces approches ont été réalisées pour un contexte segmenté avec un seul geste à la fois, et ne considèrent pas le contexte non segmenté.

La reconnaissance précoce des **3D gesture** en considérant un corps humain complet en 3D a été davantage considérée dans les travaux précédents. En ce qui concerne la reconnaissance précoce des gestes 3D, les premiers travaux ont utilisé des méthodes basées sur des modèles tentant de faire correspondre des séquences partielles de gestes [9,10]. Plus récemment, Wang et al. [11] ont conçu un modèle avec la stratégie professeur-élève, deux réseaux sont formés pour avoir une représentation interne proche. Le premier réseau (professeur) est capable de voir la séquence entière tandis que le second réseau (élève) ne peut voir qu'une partie précoce de la séquence. Wang et al. [12] ont construit un réseau capable de prédire plusieurs actions plausibles, ce qui est particulièrement utile dans les premiers stades où l'action ne peut être clairement identifiée. De plus, le modèle est entraîné avec une stratégie faiblement supervisée en prédisant les postures futures, ce qui aide le réseau à bien généraliser. Une autre façon de traiter les étapes précoces est de définir une stratégie d'option de rejet liée à la confiance comme certaines autres approches [6, 7, 13].

En ce qui concerne le cas **non segmenté**, la prédiction d'action de gestes 3D dans un flux non segmenté a été abor-

dée par Escalante et al. [14] et Liu et al. [15]. Leurs réseaux sont capables de prédire la classe à partir d'une observation partielle, mais aucune stratégie pour traiter les début de geste n'est utilisée. L'architecture de Liu et al., SSNet, est inspirée de WaveNet [16], utilisant une pile de couches convolutives 1D causales et dilatées. SSNet est capable de traiter un flux en temps réel, en donnant une nouvelle réponse à chaque nouvelle trame.

Weber et al. [2] ont abordé le problème de la reconnaissance précoce avec un réseau récurrent. Un réseau LSTM est entraîné avec une classe supplémentaire pour représenter les instants entre les gestes. Molchanov et al. [1] ont proposé un réseau récurrent convolutif (CRNN), le flux d'entrée a été divisé en petits ensembles de trames avant l'extraction de caractéristiques par un CNN. Ensuite, les caractéristiques sont introduites dans un RNN pour extraire les informations temporelles à long terme. Un système de rejet basé sur le score du classifieur est utilisé pour traiter les premiers stades des gestes. Boulahia et al. [4] ont utilisé une combinaison de SVMs entraînés avec un ensemble de caractéristiques créées à la main, un système de rejet complexe basé sur les scores de confiance a été conçu.

À partir de ces travaux, nous pouvons remarquer que la tâche de reconnaissance précoce du geste 2D n'a pas été abordée dans le contexte du flux de gestes non segmentés. Dans le contexte des gestes 3D, la plupart de ces méthodes sont entraînées avec une stratégie trame par trame, ce qui peut conduire à des résultats de prédictions instables entre des trames consécutives, rendant la méthode inutilisable dans un contexte applicatif. De plus, cette instabilité est rarement prise en compte dans la métrique d'évaluation finale puisqu'il s'agit souvent d'une métrique basée sur les résultats des trames individuelles. Dans notre travail, nous proposons une stratégie d'apprentissage qui prend explicitement en compte la stabilité temporelle avec une régularisation CTC (Connectionist Temporal Classification). De plus, nous présentons une nouvelle métrique qui pénalise durement les prédictions instables dans le temps.

3 Methode

Dans un contexte d'application où un geste est associé à une commande, chaque détection mal classifiée entraînera l'exécution d'une commande indésirable. Pour éviter de détecter quelque chose dans les premiers instants, où le geste n'est pas clairement identifiable, nous avons d'abord besoin d'un mécanisme de rejet efficace. Ensuite, nous devons assurer la stabilité des prédictions pour qu'elles soient cohérentes entre les trames consécutives.

Notre méthode étend le réseau 3D convolutif à long terme en ligne (OLT-C3D) [7] pour répondre à la tâche de reconnaissance précoce des gestes non segmentés. Tout d'abord, le signal en ligne de la trace des doigts sur le dispositif doit être traduit en une séquence d'images. Nous avons conçu deux nouvelles stratégies pour modéliser l'évolution du geste dans le temps, inspirées des stratégies précédentes. La représentation est utilisée comme entrée du réseau OLT-C3D, entraîné à traiter un flux de gestes non segmentés. Un système de rejet est utilisé pour reporter la détection dans les premières instants. Pour traiter la cohérence du geste détecté entre les trames, une sortie supplémentaire est ajoutée pour être entraînée avec la perte CTC.

3.1 Représentation du Geste

Pour choisir notre représentation des gestes, nous devons prendre en compte les deux scénarios d'application différents qui nécessitent une reconnaissance précoce dans un contexte non segmenté. Dans le premier scénario, nous considérons des gestes multi-touch effectués les uns après les autres. Entre deux tracés, tous les doigts peuvent être retirés du dispositif, et c'est également le cas entre deux gestes. Dans le second scénario, le doigt n'est jamais retiré de l'appareil, il ne fait que des gestes mono-stroke, comme lorsqu'on écrit un mot avec des lettres.

Comme la représentation dans [7], nous choisissons de représenter le geste avec une séquence d'images pour être utilisable avec le réseau OLT-C3D. À chaque nouvelle information significative nous créons une nouvelle image représentant le geste à ce stade.

Nous obtenons du dispositif le signal en ligne, c'est-à-dire une liste de points, avec l'horodatage et les positions des doigts, et nous devons le convertir en une séquence d'images. Tout d'abord, pour obtenir une représentation indépendante de la vitesse, nous rééchantillonnons le geste en utilisant la quantité de déplacement au lieu d'utiliser le temps. Entre chaque nouvelle image, la même quantité de déplacement (que nous appelons θ) a été dessinée sur le dispositif, si plusieurs traits sont effectués en même temps (geste multi-touch), alors le déplacement de tous les traits est pris en compte pour calculer la quantité de déplacement. Nous pouvons obtenir le nouvel ensemble de points S à partir de l'ensemble des points P non déjà dessinés et ordonnés dans le temps comme suit :

$$S = \left\{ p_t \in P \middle| \sum_{p_{t=1}} \|p_{t-1} - p_t\| < \theta \right\} \qquad (1)$$

Notons que cette stratégie de rééchantillonnage est applicable dans le cas en ligne et si les doigts ne bougent pas, alors aucun traitement n'est nécessaire et aucun nouveau résultat n'est donné.

Une autre difficulté est que nous ne pouvons pas deviner à l'avance quelle serait la taille du geste, l'utilisateur peut faire le geste à n'importe quelle échelle, mais notre image a une résolution spatiale finie. Pour résoudre cette difficulté, nous avons prédéfini une échelle à l'avance et si le geste atteint le bord de l'image, alors nous déplaçons l'image dans la direction opposée pour laisser de l'espace.

Pour représenter la dynamique dans l'image fixe, nous ajoutons un deuxième canal sur l'image pour notifier la présence d'un doigt sur le dispositif. Ce deuxième canal est très clairsemé et ne contient que une valeur dans les positions où se trouvent les doigts à chaque instant. Grâce

à ce canal, le réseau peut déduire dans quelle direction le trait est dessiné.

Dans l'approche précédente, le geste était traduit en une séquence d'images où chaque nouvelle image contient les nouvelles positions des doigts avec toutes ses trajectoires précédentes, faisant apparaître certains motifs. Dans un contexte non-segmenté, ceci n'est pas possible puisque nous ne savons pas quand le geste commence, et nous ne pouvons pas garder toute la trajectoire de tous les gestes car la trace se superposerait aux trajectoires des gestes précédents. Au lieu de cela, nous avons besoin d'une stratégie de représentation compatible avec une séquence de gestes. Cette stratégie doit dépendre du contexte du scénario décrit ci-dessus. Nous pouvons aborder différemment la stratégie de représentation des gestes multi-touch et celle des gestes mono-stroke.

Stratégie de Représentation Multi-Touch. Pour la représentation des gestes multi-touch, la trajectoire sera complètement réinitialisée lorsque tous les doigts auront été retirés du dispositif pendant un très court instant, qu'il s'agisse d'un moment inter-tracé ou inter-geste. De cette façon, la trajectoire du geste est accumulée jusqu'à ce que les tracés effectués simultanément soient terminés. Nous pouvons ajouter une image noire lorsque cela se produit pour notifier cet événement de manière plus explicite au réseau. De plus, cela permet au réseau de prédire quelque chose sur cette image noire tout en étant sûr que le trait est terminé, ce qui est très important pour détecter les gestes qui sont des sous-parties d'autres gestes.

Cette stratégie n'est pas applicable aux gestes mono-stroke si le doigt n'est jamais retiré de l'appareil.

Stratégie de Représentation des Gestes Mono-stroke. Pour les gestes mono-stroke, comme il n'y a pas de points de rupture identifiables, la trajectoire du geste est accumulée dans une fenêtre glissante glissante ψ. Chaque image contiendra une quantité de déplacement de $\psi \times \theta$. Une grande fenêtre de glissement conduira à des images bruitées avec une trace potentiellement superposées avec des morceaux de gestes passés, et une fenêtre trop courte ne fera apparaître aucun motif sur les dimensions spatiales. La figure 2 présente un exemple de cette représentation.

3.2 Architecture OLT-C3D avec Système de Rejet Temporel et Régularisation CTC

L'architecture OLT-C3D [7] (Online Long-Term Convolutional 3D) est composé d'un empilement de couches de convolution 3D. Les convolutions sont causales : pour calculer la sortie de chaque image, le futur est complètement ignoré. Cela garantit son utilisabilité pour les applications en temps réel. Les convolutions sont dilatées temporellement afin d'augmenter le champ réceptif dans la dimension temporelle. Avec deux blocs de 5 couches convolutionnelles avec un taux de dilatation croissant, le réseau peut faire une prédiction en voyant jusqu'à 64 images précédentes. 64 images sont suffisantes pour voir au moins un

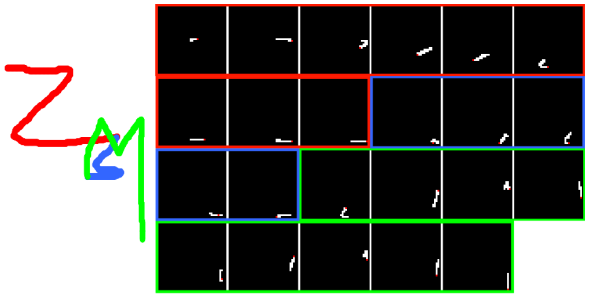


FIGURE 2 – Exemple de la représentation d'une séquence de trois gestes. Une fenêtre glissante ψ de deux unités de déplacement est utilisée dans cet exemple : chaque image contient un nouvel élément d'information avec la précédente.

geste complet. Voir [7] pour plus de détails sur l'architecture du réseau.

Nous modifions le réseau pour avoir quatre sorties : le score de confiance (1 neurone de sortie, appelé g), les scores de classification (f), la combinaison du *blank* et des scores de classe (out_{ctc}) et la sortie auxiliaire (h). Ces sorties sont représentées dans la figure 3. Trois pertes sont utilisées pour entraîner le réseau.

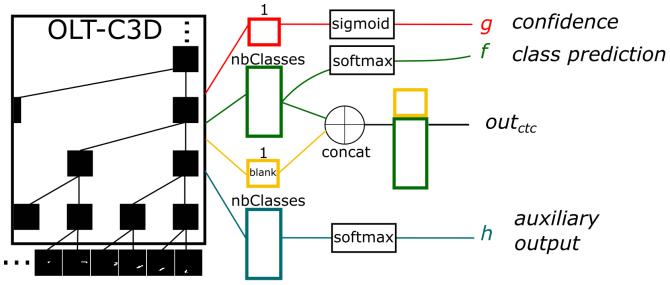


FIGURE 3 – Le réseau est composé de quatre sorties. La sortie de prédiction de classe est partagée avec la sortie CTC pour former une représentation interne commune.

3.3 Système de Rejet Temporel

Le réseau OLT-C3D est entraîné avec une perte par trame, inspirée à l'origine de la perte SelectiveNet [17], incorporant un entraînement d'un score de confiance.

Cette perte est calculée comme suit :

$$\mathcal{L}_{(f,g)} = \frac{1}{m}\sum_{i=1}^{m} \ell\left(f\left(x_i\right), y_i\right) g\left(x_i\right) + \lambda \Psi\left(c - \hat{\phi}\left(g\right)\right) \tag{2}$$

où $\Psi(a) = \max(0, a)^2$, λ est un hyperparamètre relatif à l'importance de la contrainte de couverture (nous le fixons à $\lambda = 32$ comme l'approche précédente). c est la couverture ciblée. $\hat{\phi}\left(g\right)$ est la couverture empirique, c'est-à-dire la valeur moyenne de $g(x)$, et ℓ est la perte d'entropie croisée.

Pendant l'inférence, nous considérerons que la trame de

prédiction est acceptée si g est supérieur à un seuil γ, fixé à 0,5.
La sortie auxiliaire h est constituée de prédictions de classe comme f, mais est entraînée par une perte d'entropie croisée traditionnelle par trame $\mathcal{L}_h$.
La perte de rejet temporel et la perte auxiliaire sont des pertes par trame, et aucune considération n'est donnée à la cohérence entre trames consécutives. Nous abordons ce problème avec la régularisation CTC.

3.4 Régularisation avec la Perte de la Classification Temporelle Connexionniste (CTC)

La sortie out_{ctc} du réseau est entraînée avec la perte CTC $\mathcal{L}_{ctc}$. Notez que les scores de classe qui font partie de out_{ctc} sont partagés avec la perte de rejet temporel.
La perte CTC [18] est utilisée pour optimiser l'alignement entre une annotation au niveau de la séquence (c'est-à-dire les classes se produisant pendant la séquence, sans les limites de segmentation de début/fin) et une sortie par trame. La sortie par trame est traitée dans la perte CTC pour supprimer les prédictions identiques consécutives de la sortie par image. Un *blank* est également utilisé comme un rejet pour ne prédire aucune des classes possibles.
L'utilisation de la perte CTC n'est pas nécessaire pour l'entraînement du réseau puisque nous avons la segmentation temporelle des gestes, nous aurions pu entraîner notre réseau uniquement avec la perte par trame. Mais la perte CTC apporte une stabilité de prédiction temporelle au réseau car elle nécessite un bon alignement entre l'annotation de niveau séquence et la séquence prédite, ce qui est très utile pour notre tâche. Nous pouvons remarquer que le rôle du *blank* dans la CTC est relativement proche de celui du score de confiance du système de rejet temporel. Le principal avantage du score de confiance par rapport au blanc est que nous pouvons facilement le modifier en utilisant différents paramètres (couverture ciblée c, λ, seuil de confiance γ).
Pour la détection finale, nous utilisons la sortie de prédiction de classe f avec le score de confiance g. La perte CTC est juste utilisée pour entraîner la représentation interne du réseau et pour lisser la prédiction de classe au fil du temps, nous pouvons la considérer comme une stratégie de régularisation.

3.5 Perte finale optimisée

La perte finale optimisée est calculée comme suit :

$$\mathcal{L} = \alpha\mathcal{L}_{(f,g)} + (1-\alpha)\mathcal{L}_h + \omega\mathcal{L}_{ctc} \quad (3)$$

où nous avons fixé $\alpha = 0,5$ et $\omega = 0,01$ pour que l'ordre de grandeur de la perte CTC soit proche de celle des autres pertes.

3.6 Détails du réseau

Les images générées par notre représentation sont de 40 par 40 pixels, avec un déplacement de quantité θ égal à 4,4 pour ILGDB et 1,5 pour MTGSetB (une fois mis à l'échelle par 0,2 pour ILGDB et 0,03 pour MTGSetB). La fenêtre glissante ψ pour la représentation de l'ILGDB est fixée à 2. Le Dropout est utilisé dans toutes les couches convolutionnelles et denses, avec un taux de 0,1 pour les couches convolutionnelles et de 0,2 pour les couches denses. Chaque couche convolutive apprend 30 filtres. Une couche dense de 100 unités est utilisée après les couches convolutionnelles, toutes les sorties partagent cette couche. Une couche dense supplémentaire, avec le même nombre d'unités, est utilisée juste avant la couche de sortie de confiance finale. Pendant l'apprentissage, une rotation aléatoire (suivant une distribution normale avec $\mu = 0$ et $\sigma = 15°$) est appliquée à la séquence (la même rotation pour toutes les images d'une séquence) pour améliorer la généralisation. La couverture c de la perte par rejet temporel est fixée à 0,7 pour MTGSetB et à 0,3 pour ILGDB. L'entraînement est effectué avec une taille de *batch* de 5 séquences.

4 Expériences

4.1 Génération de jeux de données synthétiques

Pour évaluer notre approche sur cette tâche, nous avons généré deux jeux de données à partir de ILGDB [19] et MTGSetB [20]. ILGDB est un jeu de données de gestes mono-stroke contenant 21 classes de gestes effectués par 38 utilisateurs. Ces 21 classes sont divisées en 7 groupes de 3 classes, où ces 3 classes partagent un début commun, rendant presque impossible la détection précoce avant la bifurcation de ces 3 gestes. Nous générons des séquences avec entre 4 et 8 gestes sélectionnés aléatoirement. La séquence est générée de manière à ce que le dernier point d'un geste soit le même que le premier point du geste suivant. En suivant la répartition originale train/test, 119 séquences sont utilisées pour l'apprentissage et 210 pour le test. Ce jeu de données est particulièrement difficile car les séquences ne comportent aucune pause et il est donc très difficile de déterminer le début et la fin des gestes. De plus, il comporte peu d'exemples d'entraînement. Pour faire face à la faible quantité de données, nous avons généré un jeu de données augmenté en utilisant différentes échelles de gestes (5 échelles différentes) et en utilisant le même geste dans plusieurs séquences (chaque geste est mis dans 5 séquences). Au final, chaque geste d'entraînement est utilisé 25 fois dans les séquences, ce qui donne 2621 séquences d'entraînement. Un exemple de séquence générée est donné dans la figure 1.
MTGSetB est un ensemble de données de gestes multi-touch contenant 31 classes différentes effectuées par 33 utilisateurs. Nous avons construit des séquences de 4 à 8 gestes aléatoires de sorte à ne pas être en mesure de différencier un blanc entre les gestes et un blanc entre les gestes, chaque geste étant recentré en fonction du geste précédent dans la séquence, rendant impossible une segmenta-

tion spatiale entre les gestes. Selon la répartition originale entre l'entraînement et le test, séparée par utilisateurs, on obtient 607 séquences de gestes pour l'entraînement et 672 pour le test. Nous avons également généré une version augmentée de l'ensemble de données avec différentes échelles (3 échelles différentes) et en utilisant chaque geste dans 2 séquences. Nous avons ainsi obtenu 3076 séquences d'entraînement.
Ces deux jeux de données sont disponibles librement [1]

4.2 Métrique de détection en ligne bornée (BOD)

Nous proposons une nouvelle métrique appelée *"Bounded Online Detection (BOD) Metric"*, inspirée des métriques de détection en vision par ordinateur. L'idée principale de cette métrique est de n'autoriser qu'une seule détection par bornes de vérité terrain, en conditionnant la détection à une certaine quantité de chevauchement entre la borne de la vérité terrain et la borne de détection. Toute détection qui n'a pas assez de chevauchement ou qui n'a pas la bonne classification sera considérée comme un faux positif.
Notez que pour calculer cette métrique, nous avons besoin des bornes des prédictions. Si nous avons une sortie de classification par trame, nous aurons besoin d'une stratégie pour la transformer en une sortie bornée. Cette stratégie doit être compatible avec le contexte en ligne, c'est-à-dire que nous ne devons pas utiliser les prédictions futures pour estimer une borne de début et de fin. Dans notre cas, la première image acceptée est considérée comme le début du geste, et le rejet suivant ou la prédiction d'une classe différente est la fin de ce geste.
Nous avons conçu le IoU_{st} qui est une variante de la mesure *Intersection Over Union* pour le contexte en ligne. Comme nous ne voulons pas pénaliser la prédiction tardive sur ce critère, nous calculons le chevauchement à partir de la borne de début de prédiction. Un IoU_{st} élevé caractérise une détection qui correspond bien à la born de la vérité de terrain à partir de la borne de début de prédiction.
La métrique a deux paramètres : "*canCorrect* (*CC*)" et Δ. *canCorrect* est un booléen, s'il est vrai, il donne au modèle la possibilité de se corriger s'il a fait une erreur de détection sur un geste donné. Δ est une valeur comprise entre 0 et 1, et correspond à la valeur minimale de IoU_{st} autorisée pour considérer le geste comme un vrai positif. Pour un contexte applicatif qui ne nécessite pas de conserver la prédiction pendant le geste (juste un pic), Δ peut être fixé à 0. Une valeur de 1 signifierait que la borne de fin de la prédiction devrait correspondre exactement à la borne de fin de la vérité de terrain pour être un vrai positif.
La détection est considérée comme un *vrai positif* (*TP*) si la vérité terrain avec le IoU_{st} maximum n'a pas déjà été correctement détectée (ou faussement détectée si $canCorrect = False$), s'il s'agit de la prédiction de classe correcte, et si le IoU_{st} est strictement supérieur à Δ. Sinon, elle est considérée comme un *false positive* (*FP*). Un exemple d'application de la métrique est présenté dans la figure 4.

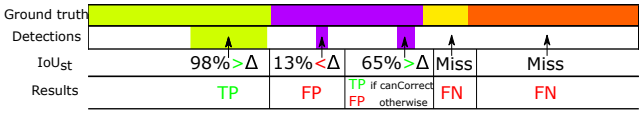


FIGURE 4 – Exemple d'une application de la métrique BOD (Bounded Online Detection) avec $\Delta = 50\%$.

Une fois la *Precision* et le *Rappel* calculés, nous pouvons calculer la la *FMesure* globale finale en utilisant le calcul traditionnel : $FMesure = \frac{2*Rappel*Precision}{Precision+Rappel}$, et en calculant ensuite sa micro-moyenne. Nous calculons également la distance normalisée de détection (NDToD) utilisée dans des travaux antérieurs [6, 7], qui mesure la précocité de la détection. Cette métrique ne prend en compte que les détections *positives*.

4.3 Résultats

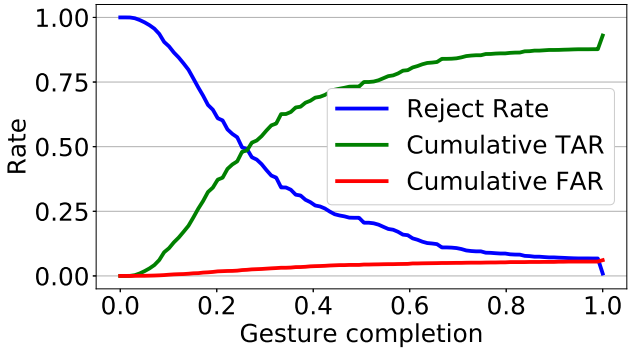
Résultats de la reconnaissance précoce en contexte segmenté. Pour comparer notre approche aux travaux précédents, nous l'avons évaluée dans un contexte segmenté. Dans ce contexte, seule la première détection est utilisée car une application saurait qu'un seul geste est dessiné. La comparaison effectuée avec le jeu de données MTGSetB est présentée dans le tableau 1. Notre méthode a amélioré de manière significative les valeurs taux d'acceptation positif (TAR) et taux d'acceptation négatif (FAR). Presque aucun geste n'est rejeté (RR : Taux de rejet). Néanmoins, les détections se produisent un peu plus tard que dans le cas de [7]. Cela montre en particulier le bon impact de la régularisation CTC, même dans un contexte segmenté.

TABLE 1 – Comparaison avec les approches précédentes pour une valeur de précocité équivalente (NDToD), évaluation en segmenté

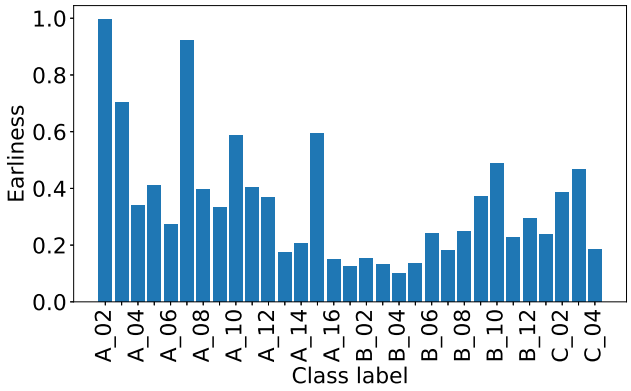
Method	TAR	FAR	RR	NDtoD
Chen et al. [6]	81.89 %	14.56 %	3.54 %	37.04 %
OLT-C3D [7]	89.25 %	7.24 %	3.51 %	**30.77** %
This work	**93.5** %	**6.44** %	**0.1** %	33.1 %

La figure 5a montre le comportement du système selon l'évolution du geste. Nous pouvons voir que 50 % des gestes ont été détectés avant 27 % de leur achèvement. Parmi eux, 95,4 % ont été correctement classés. Comme le montre la figure 5b, la précocité peut varier considérablement entre les catégories de gestes, de près de 100 % pour A_02 à moins de 15 % pour B_04. Cela dépend notamment de la longueur du début commun entre les gestes. En ce qui concerne l'exactitude (*accuracy*) sans précocité, en prenant la prédiction à la dernière image pour chaque geste, nous obtenons un score de 97,33 % contre 94,45 % pour [7].

1. Jeux de données disponibles sur : https://www-intuidoc.irisa.fr/en/mtgsetb-and-ilgdb-untrimmed/



(a) Comportement du système en ce qui concerne le taux d'acceptation positif (TAR), le taux d'acceptation négatif (FAR) et le taux de rejet (RR) en fonction du taux d'achèvement du geste.



(b) Précocité par classe.

FIGURE 5 – Comportement et précocité du système sur le jeu de données MTGSetB, évaluation en segmenté.

4.4 Résultats de la reconnaissance précoce non segmenté

Nous avons évalué notre approche dans le contexte non segmenté sur les deux ensembles de données précédemment décrits avec la métrique BOD. Les résultats sont présentés dans le tableau 2. La méthode a été évaluée avec différentes combinaisons de paramètres de la métrique. Pour MTGSetB, nous avons obtenu une FMesure de 83,6 % à 76,1 % selon la valeur du Δ, sans possibilité de corriger l'erreur, et de 88,2 % à 86,8 % si le modèle peut se corriger. Ces résultats montrent que le réseau est capable de bien prédire jusqu'à la fin du geste pour ce jeu de données. Pour ILGDB, les résultats varient beaucoup plus en fonction de Δ, cela signifie que le réseau n'est pas capable d'identifier correctement la fin du geste, cela est cohérent avec la difficulté de ce jeu de données car les gestes sont totalement enchaînés, et les transitions entre les gestes ne sont pas facilement identifiables.

TABLE 2 – Évaluation en non-segmenté sur MTGSetB et ILGDB avec different parametres de la métrique BOD. CC : CanCorrect.

		MTGSetB		**ILGDB**	
CC	Δ	FMeasure	NDToD	FMeasure	NDToD
False	0.0	83.6 %	32.7 %	61.1 %	68.7 %
	0.5	77.1 %	32.5 %	45.1 %	68.2 %
	0.95	76.1 %	32.4 %	24.3 %	71.9 %
True	0.0	88.2 %	34.0 %	68.0 %	69.3 %
	0.5	87.7 %	35.8 %	54.3 %	70.4 %
	0.95	86.8 %	36.1 %	30.9 %	75.4 %

4.5 Impact de la régularisation CTC

Pour montrer l'importance de la régularisation CTC, nous avons évalué le système avec et sans cette régularisation, les résultats sont présentés dans le tableau 3. Sur les deux jeux de données, la régularisation CTC a un impact significatif sur les résultats, en particulier sur ILGDB. En raison de la perte CTC qui encourage les prédictions à être cohérentes entre les images, la précision et la précocité sont particulièrement impactées : +1,6 % de précision pour MTGSetB et +8,8 % pour ILGDB. Nous pouvons en déduire que le réseau a préféré reporter davantage sa détection pour éviter l'instabilité de la détection.

Résultats qualitatifs. La figure 6 (en haut) montre un exemple d'une séquence de MTGSetB avec les prédictions et les détections du réseau. Pour le premier geste, on voit que le réseau rejette les prédictions (i.e., confiance inférieure à 0.5) jusqu'au début du deuxième tracé du geste "X" pour être sûr de ne pas confondre avec le geste "W" qui est aussi contenu dans le jeu de données. Pour les deux gestes suivants, il attend également que les parties communes avec d'autres gestes soient passées. Nous observons un comportement similaire sur ILGDB (figure 6, en bas), mais en raison de la difficulté de ce jeu de données, il a du mal à conserver un score de confiance cohérent, ce qui peut produire des faux positifs et des détections manquées.

5 Conclusion

Nous avons présenté une approche visant à relever le défi de la reconnaissance précoce des gestes non segmentés. Tout d'abord, de nouvelles stratégies de représentation sont conçues pour prendre en compte des séquences de gestes à un ou plusieurs tracés. Nous avons proposé de régulariser le CNN spatio-temporel en utilisant la perte CTC pour apporter une stabilité aux prédictions. De plus, nous proposons un nouveau protocole d'évaluation avec une nouvelle métrique, et deux jeux de données artificiels. Notre méthode a obtenu des résultats supérieurs à ceux d'autres approches en contexte segmenté. De bons résultats sont obtenus dans un contexte non segmenté, ce qui constituera un score de référence solide pour les futurs travaux.
L'application de la perte CTC peut être une porte ouverte à l'apprentissage faiblement supervisée, en utilisant uniquement l'annotation au niveau de la séquence, ce qui sera ex-

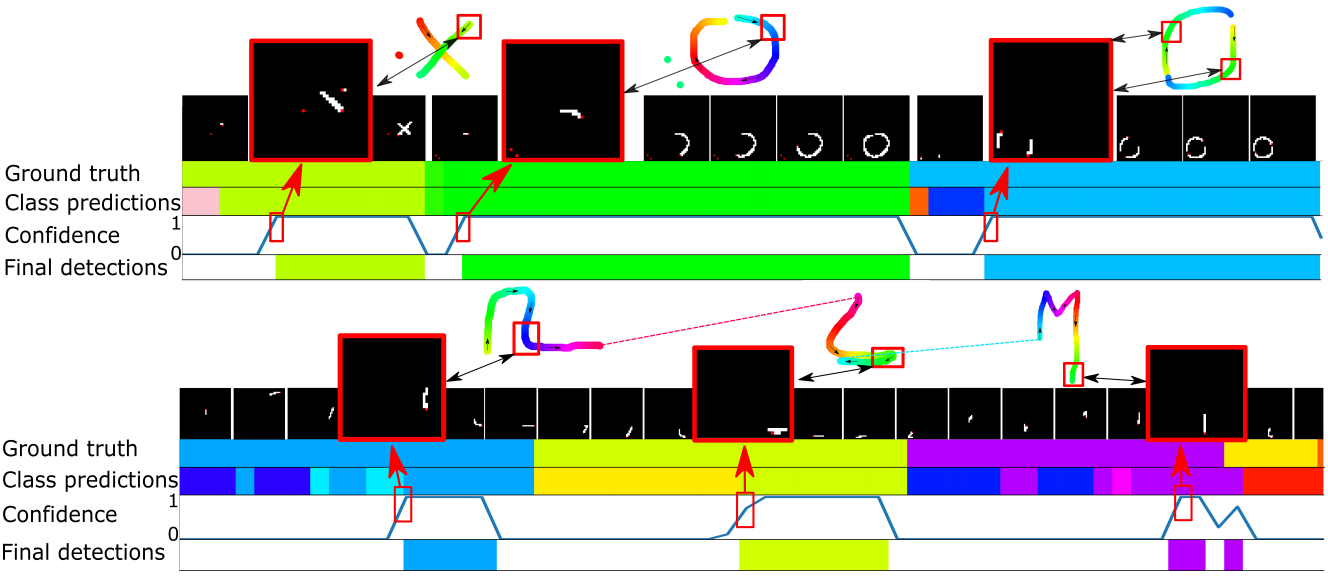


FIGURE 6 – Exemple de détections sur MTGSetB (en haut) et ILGDB (en bas). Le réseau attend les instants décisifs pour obtenir une confiance élevée. Notez que certaines images intermédiaires ont été supprimées pour la visibilité.

TABLE 3 – Impact de la regularisation avec le CTC, contexte non segmenté, métrique BOD avec $canCorrect = False, \Delta = 0.25$

Dataset	Variant	Precision	Recall	FMeasure	NDtoD
MTGSetB	CTC	**70.7 %**	**87.0 %**	**78.0 %**	32.7 %
	No CTC	69.1 %	85.2 %	76.3 %	**32.3** %
ILGDB	CTC	**56.4 %**	**55.6** %	**56.0** %	69.2 %
	No CTC	47.6 %	54.2 %	50.7 %	**66.5 %**

ploré dans les travaux futurs. Nous nous pencherons également sur la tâche de reconnaissance précoce d'actions d'un corps humain en 3D.

Remerciement

Cette étude est financée par l'ANR dans le cadre du projet PIA EUR DIGISPORT (ANR-18-EURE-0022).

Références

[1] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[2] M. Weber, M. Liwicki, D. Stricker, C. Scholzel, and S. Uchida, "Lstm-based early recognition of motion patterns," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 3552–3557.

[3] V. Bloom, V. Argyriou, and D. Makris, "Linear latent low dimensional space for online early action recognition and prediction," *Pattern Recognition*, vol. 72, pp. 532–547, 2017.

[4] S. Y. Boulahia, E. Anquetil, F. Multon, and R. Kulpa, "Détection précoce d'actions squelettiques 3D dans un flot non segmenté à base de modèles curvilignes," in *RFIAP 2018 Reconnaissance des Formes, Image, Apprentissage et Perception*, Paris, France, Jun. 2018, pp. 1–8.

[5] S. Uchida and K. Amamoto, "Early recognition of sequential patterns by classifier combination," in *19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[6] Z. Chen, E. Anquetil, C. Viard-Gaudin, and H. Mouchère, "Early recognition of handwritten gestures based on multi-classifier reject option," in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017, pp. 212–217.

[7] W. Mocaër, E. Anquetil, and R. Kulpa, "Online spatio-temporal 3d convolutional neural network for early recognition of handwritten gestures," in *Document Analysis and Recognition – ICDAR 2021*, J. Lladós, D. Lopresti, and S. Uchida, Eds. Cham : Springer International Publishing, 2021, pp. 221–236.

[8] M. Yamagata, H. Hayashi, and S. Uchida, "Handwriting prediction considering inter-class bifurcation structures," in *17th International Conference*

on Frontiers in Handwriting Recognition (ICFHR), 2020, pp. 103–108.

[9] A. Mori, S. Uchida, R. Kurazume, R. Taniguchi, T. Hasegawa, and H. Sakoe, "Early recognition and prediction of gestures," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3, 2006, pp. 560–563.

[10] M. Kawashima, A. Shimada, H. Nagahara, and R. Taniguchi, "Adaptive template method for early recognition of gestures," in *17th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*, 2011, pp. 1–6.

[11] X. Wang, J.-F. Hu, J.-H. Lai, J. Zhang, and W.-S. Zheng, "Progressive teacher-student learning for early action prediction," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3551–3560.

[12] R. Wang, J. Liu, Q. Ke, D. Peng, and Y. Lei, "Dearnet : Learning diversities for skeleton-based early action recognition," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[13] S. Y. Boulahia, E. Anquetil, F. Multon, and R. Kulpa, "Cudi3d : Curvilinear displacement based approach for online 3d action detection," *Computer Vision and Image Understanding*, vol. 174, pp. 57 – 69, 2018.

[14] H. J. Escalante, E. F. Morales, and L. E. Sucar, "A naïve bayes baseline for early gesture recognition," *Pattern Recognition Letters*, vol. 73, pp. 91 – 99, 2016.

[15] J. Liu, A. Shahroudy, G. Wang, L. Duan, and A. C. Kot, "Skeleton-based online action prediction using scale selection network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 6, pp. 1453–1467, 2020.

[16] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet : A generative model for raw audio," *CoRR*, 2016.

[17] Y. Geifman and R. El-Yaniv, "Selectivenet : A deep neural network with an integrated reject option," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2151–2159.

[18] A. Graves, S. Fernández, and F. Gomez, "Connectionist temporal classification : Labelling unsegmented sequence data with recurrent neural networks," in *In Proceedings of the International Conference on Machine Learning, ICML 2006*, 2006, pp. 369–376.

[19] N. Renau-Ferrer, P. Li, A. Delaye, and E. Anquetil, "The ilgdb database of realistic pen-based gestural commands," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, pp. 3741–3744.

[20] Z. Chen, E. Anquetil, H. Mouchère, and C. Viard-Gaudin, "Recognize multi-touch gestures by graph modeling and matching," in *17th Biennial Conference of the International Graphonomics Society*, ser. Drawing, Handwriting Processing Analysis : New Advances and Challenges. Pointe-a-Pitre, Guadeloupe : International Graphonomics Society (IGS) and Université des Antilles (UA), Jun. 2015.