Approche hybride pour la détection du levé de stylo à partir de vidéos : preuve de concept pour une analyse complémentaire de l'écriture

Lauren Sismeiro[1], Rémy Plastre[2], Binbin Xu[1], Frédéric Puyjarinet[1], Gérard Dray[1]

[1] EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Alès, France
[2] IMT Mines Alès, Alès, France

lauren.sismeiro@mines-ales.fr, remy.plastre@mines-ales.org

Résumé

L'analyse dynamique de l'écriture est essentielle pour l'évaluation de la dysgraphie, mais les tablettes graphiques ne capturent que les mouvements proches de la surface. Nous proposons une preuve de concept visant à détecter les levés de stylo à partir de vidéos en vue de dessus. L'approche repose sur une chaîne de traitement hybride combinant suivi de pointe du stylo, descripteurs cinématiques multi-échelles et classification supervisée. Évaluée en Leave-One-Video-Out, la méthode atteint un score F_2 de 0.805, suggérant que l'analyse vidéo constitue un complément pertinent et peu coûteux aux dispositifs existants.

Mots-clés

Analyse de l'écriture manuscrite, Vision par ordinateur, Suivi de stylo, Cinématique, Dysgraphie, Santé numérique.

1 Introduction

Le diagnostic de la dysgraphie repose notamment sur l'échelle BHK [1], fondée sur l'analyse statique de l'écriture, sans accès à la dynamique du geste, pourtant essentielle pour caractériser le contrôle moteur. Les tablettes numériques capturent ces informations, mais restent limitées aux mouvements proches de la surface [2], excluant les levées de grande amplitude potentiellement informatives [4]. Nous proposons d'explorer la vision par ordinateur comme modalité complémentaire. À partir de vidéos en vue de dessus, nous évaluons dans une approche de preuve de concept, la capacité d'une chaîne de traitement hybride à détecter les états de contact (*Pen-Down*) et de levé (*Pen-Up*).

2 Méthode

2.1 Collecte de données et annotation

Cinq vidéos d'écriture manuscrite ont été extraites de YouTube[1], couvrant divers styles (cursive, script), stylos (pointes fines/épaisses, encres bleue/noire) et supports (interlignes variables). Le jeu de données ainsi créé comprenait 13 507 images après échantillonnage à 30 fps, d'une résolution de 1080p. Chaque frame a été annotée de façon binaire par 3 évaluateurs indépendants en *Pen-Down* ou *Pen-Up*, en excluant les images non informatives, atteignant un accord inter-juges de $89\,\%$ et un coefficient Kappa de Fleiss de $0,78$. Après agrégation via soft labelling, on comptait $28,7\,\%$ d'images étiquetées *Pen-Up*.

1. DorufaVSArt, `https://www.youtube.com/@DorufaVSArt`

2.2 Métrique d'évaluation

Le score F_2 dans la détection des évènements *Pen-Up* a été choisi comme la métrique prioritaire, afin de privilégier des modèles sensibles avec un rappel élevé plus adaptés en contexte de dépistage.

2.3 Chaîne de traitement en quatre étapes

L'approche proposée, illustrée Figure 1, repose sur une chaîne de traitement hybride séparant explicitement la localisation de la pointe du stylo et l'inférence de son état.

1. Suivi de la pointe du stylo. La position de la pointe (u, v) a été estimée image par image à l'aide du modèle de détection d'objets YOLOv11m, selon un protocole Leave-One-Video-Out. Le modèle a atteint une erreur médiane de 3,28 px (P95 : 7,44 px), avec 77,7 % des prédictions < 5 px et 99,2 % < 10 px, permettant d'obtenir une trajectoire fiable, y compris lors de mouvements rapides ou de levés importants comme l'illustre la Figure 2.

2. Extraction de caractéristiques cinématiques. À partir des coordonnées (u, v), 147 descripteurs cinématiques ont été extraits, comprenant notamment des caractéristiques locales multi-échelles (fenêtres glissante de 3 à 16 images) de linéarité, de variabilité angulaire et de variations de vitesse, ainsi que des descripteurs globaux (inclinaison moyenne, vitesse normalisée) tenant compte du style d'écriture.

3. Classification supervisée. Random Forest, HistGBM, LightGBM ainsi qu'une approche par Ridge stacking ont été entraînés et évalués selon le protocole Leave-One-Video-Out, afin de maximiser la F_2 des probabilités $P_{\text{Pen-Up}}$ par image, incluant une optimisation des hyperparamètres via la librairie Optuna (totalisant 100 essais par modèle).

4. Post-traitement événementiel. Les probabilités ont été converties en segments temporels (évènements *Pen-Up* et *Pen-Down*) grâce à un post-traitement combinant : **(i)** un lissage par hystérésis pour stabiliser les transitions, **(ii)** un filtrage morphologique pour supprimer les détections bruitées, et **(iii)** un alignement cinématique.

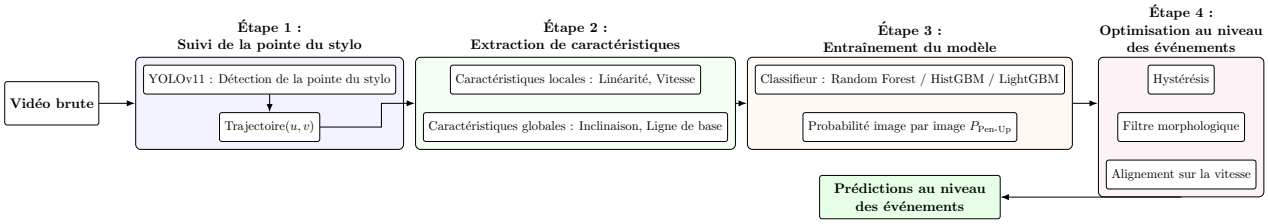

FIGURE 1 – Chaîne de traitement hybride pour la détection du contact du stylo en vidéos d'écriture manuscrite vues de dessus.

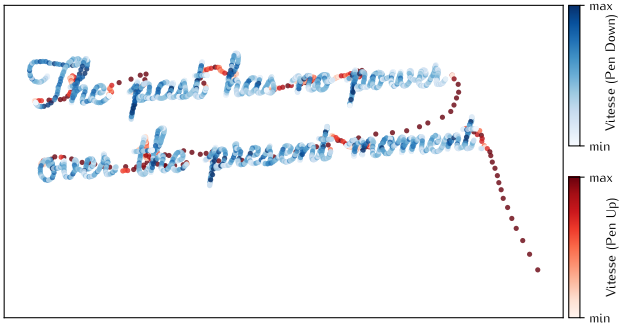

FIGURE 2 – Visualisation de la trajectoire du stylo détectée par Yolov11m (Bleu : *Pen-Down*, Rouge : *Pen-Up*, vitesse codée en intensité)

3 Résultats

L'évaluation a été réalisée au niveau événementiel, afin de détecter des segments de levé de stylo plutôt que des prédictions image par image. Plusieurs tolérances temporelles de décalage entre prédiction et annotation ont ainsi été testées (0 à 20 frames). Les meilleures performances ont été obtenues via les modèles de type gradient boosting. À une tolérance de 12 images ($\approx$ 400 ms, durée considérée comme cliniquement pertinente pour les pauses en écriture [3]), le modèle LightGBM a atteint un score F_2 de $0,805$ avec un rappel de $0,880$. La Figure 3 illustre l'évolution du rappel et du score F_2 en fonction de la tolérance temporelle. L'augmentation de cette tolérance améliorait les performances, en particulier à faible valeur (< 5 frames). Les gains observés entre 10 et 12 images étaient quant à eux limités, suggérant un plateau de performance.

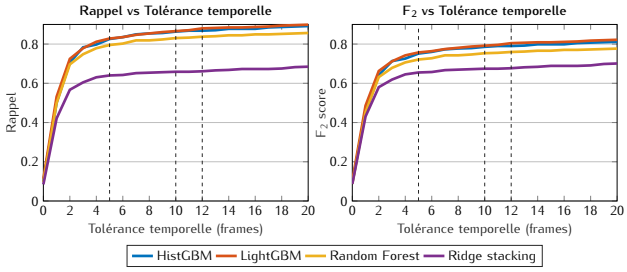

FIGURE 3 – Rappel et Score F_2 en fonction de la tolérance pour les quatre modèles évalués.

4 Discussion

Cette étude démontre la faisabilité de la détection des états *Pen-Up* à partir de vidéos en vue de dessus, comme complément possible aux tablettes numériques. La chaîne de traitement proposée repose sur des descripteurs cinématiques interprétables, facilitant ainsi l'analyse et l'usage en contexte clinique. Ces mesures pourraient notamment fournir aux psychomotriciens des indicateurs objectifs du geste.

Limites et perspectives. Le jeu de données reste de taille réduite et devra être étendu pour confirmer la robustesse des résultats. Par ailleurs, l'utilisation d'une caméra monoculaire ne permet pas d'estimer directement la hauteur du stylo, les états *Pen-Up* étant inférés indirectement à partir de la cinématique. Des approches multi-vues pourraient permettre d'accéder à cette information.

5 Conclusion

Ce travail constitue une première preuve de concept montrant qu'une analyse vidéo peut compléter les tablettes graphiques pour l'étude de l'écriture manuscrite. En permettant le suivi de la trajectoire au-delà de la surface d'écriture, cette approche ouvre la voie à une caractérisation plus complète des dynamiques scripturales, avec des perspectives d'application en évaluation clinique de la dysgraphie.

Remerciements

Les auteurs remercient le créateur DorufaVSArt pour l'autorisation d'utilisation des vidéos, ainsi que Romain Sebire pour sa contribution à l'annotation des images.

Références

[1] M. Charles, R. Soppelsa, and J.-M. Albaret. *BHK – Échelle d'évaluation rapide de l'écriture chez l'enfant*. Éditions Centre de Psychologie Appliquée, 2004.

[2] Jean-Claude Gilhodes, Elie Fabiani, Marieke Longcamp, Jean-Luc Velay, and Jérémy Danna. Chapitre 4. Traiter des données de langage écrit recueillies avec tablette graphique. In *Introduction aux statistiques en sciences du langage*, pages 117–136. Dunod, 2023.

[3] Mariona Pascual, Olga Soler, and Naymé Salas. In a split second : Handwriting pauses in typical and struggling writers. *Frontiers in Psychology*, Volume 13 - 2022, 2023.

[4] Sara Rosenblum, Shula Parush, and Patrice L. Weiss. The in Air Phenomenon : Temporal and Spatial Correlates of the Handwriting Process. *Perceptual and Motor Skills*, 96(3) :933–954, June 2003.