

Reconnaissance égocentrique de langue des signes : De la vidéo à la synthèse vocale

2026 – 2027

Rémi Giraud

remi.giraud@enseirb-matmeca.fr

Contexte

Très peu de personnes à travers le monde sont formées à la langue des signes, qui repose notamment sur la configuration des mains, leurs mouvements et leur évolution temporelle. La reconnaissance automatique de ce langage est donc un problème intéressant de vision par ordinateur et d'utilité publique.

Des bases de données telles que WLASL [1] permettent d'étudier la reconnaissance de signes à partir de séquences vidéo, souvent filmées face à la personne. Une approche particulièrement intéressante mais très sous exploitée est la *vision égocentrique*, dans laquelle la caméra est portée directement par la personne réalisant les signes. Un tel système de reconnaissance permettrait de synthétiser en temps-réel les signes exécutés soi-même. Cependant le changement de point de vue, les occultations et les mouvements de caméra rendent la reconnaissance plus difficile.

Objectif

L'objectif de ce projet est de mettre en œuvre une **chaîne complète de traitement en temps réel**, depuis l'acquisition vidéo jusqu'à la reconnaissance du signe et sa restitution sous forme de texte et de parole.

Le système sera d'abord développé à partir de WLASL [1], puis adapté à des vidéos acquises en **vue égocentrique** à l'aide d'une petite caméra USB. Cette étape permettra notamment d'étudier l'influence du changement de point de vue sur les représentations utilisées et les performances de reconnaissance.

Enfin, la chaîne pourra être déployée sur un **Raspberry Pi** afin de réaliser un démonstrateur autonome.

Chaîne de traitement envisagée

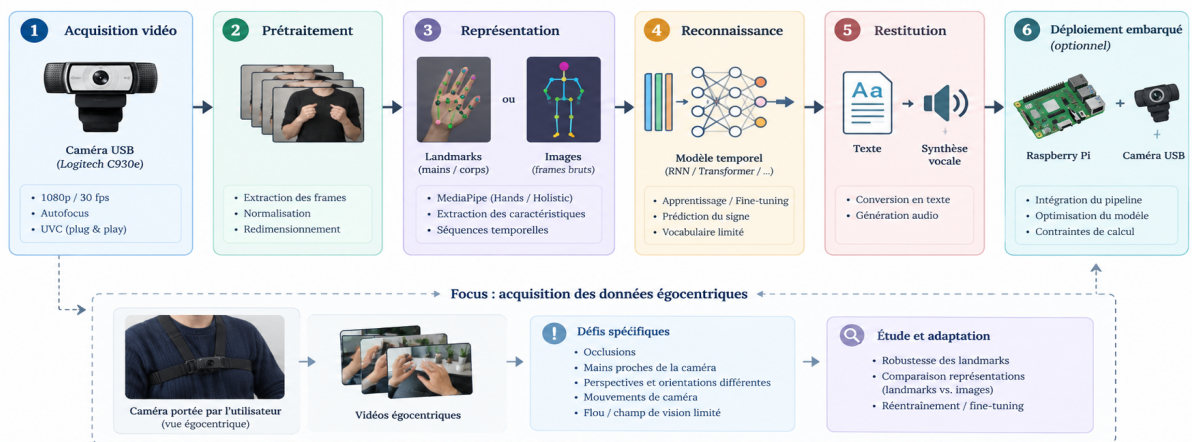


FIGURE 1 – Chaîne de traitement envisagée : de l'acquisition vidéo à la synthèse vocale et au déploiement embarqué.

Étapes du projet

Certaines de ces étapes pourront être menées en parallèle.

1. État de l'art et prise en main

- Étude des méthodes de reconnaissance de langue des signes et prise en main de WLASL.
- Mise en œuvre d'un premier modèle sur un vocabulaire limité.

2. Chaîne de reconnaissance temps réel

- Étude de représentations basées images et les *landmarks* des mains et du corps [2].
- Mise en œuvre de la chaîne vidéo → reconnaissance → texte → synthèse vocale.

3. Passage à la vision égocentrique

- Acquisition d'un petit jeu de données avec une caméra USB portée par l'utilisateur.
- Étude de la robustesse des *landmarks* et adaptation éventuelle du modèle par réentraînement ou *fine-tuning*.

4. Déploiement embarqué

- Intégration de la caméra et de la chaîne de traitement sur un Raspberry Pi.
- Évaluation des performances et des contraintes de calcul.

Résultats et rendu

Le projet devra aboutir à une **chaîne complète de démonstration** permettant de reconnaître un signe réalisé en vue égocentrique et de restituer le résultat sous forme de texte/parole.

Le système sera évalué en termes de taux de reconnaissance, de temps de traitement et de robustesse aux variations entre utilisateurs. Lorsque les performances le permettent, la chaîne sera déployée sur un **Raspberry Pi associé à une caméra USB** afin de réaliser un démonstrateur autonome en temps réel.

Le rendu comprendra également les éléments nécessaires à la présentation du travail : support de présentation, poster, vidéo de démonstration ou autre support pertinent, ainsi qu'une analyse des choix techniques et des résultats.

Références

- [1] D. Li, C. R. Opazo, X. Yu, and H. Li.
Word-level Deep Sign Language Recognition from Video : A New Large-scale Dataset and Methods Comparison.
Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2020.
<https://github.com/dxli94/WLASL>.
- [2] F. Zhang et al.
MediaPipe Hands : On-device Real-time Hand Tracking.
<https://github.com/google-ai-edge/mediapipe>.
- [3] A. Duarte et al.
How2Sign : A Large-scale Multimodal Dataset for Continuous American Sign Language.
<https://how2sign.github.io/>.