Analyse de vidéos de Langue des Signes Française à des fins de reconnaissance automatique

Valentin Belissen
LIMSI, CNRS,
Université Paris-Saclay

Annelies Braffort
LIMSI, CNRS,
Université Paris-Saclay

Michèle Gouiffès
LIMSI, CNRS,
Université Paris-Saclay

La problématique de recherche porte sur l'analyse automatique de vidéos de Langue des Signes Française (LSF). Les vidéos traitées sont issues de sources très diverses, et dans lesquelles les locuteurs ne portent pas de capteurs. Le champ d'application est donc large. Ce projet, comme détaillé ci-dessous, comporte deux volets principaux : le premier concerne le traitement de l'image, le second l'analyse des informations extraites de l'image à des fins de traitement automatique de la langue.

Les premiers travaux de ce projet concernent la réduction de la dimensionnalité de l'information présente dans une vidéo de LSF. En effet, dans l'optique d'effectuer un traitement linguistique approprié, il est nécessaire d'extraire de chaque image composant la vidéo les seules informations pertinentes, c'est-à-dire de suivre correctement les articulateurs du buste, de la tête, des bras, des mains et du visage. Il a été décidé de travailler séparément sur : le squelette associé à la posture ; le visage ; les mains.

Grâce à des outils comme OpenPose [1] qui utilisent des réseaux de neurones convolutionnels, la posture est récupérée de manière assez précise, en deux dimensions. Cependant, la nature tridimensionnelle de la LSF nous a incités à entraîner un second réseau de neurones profond [2] avec les données du corpus Mocap1 [3], où 8 locuteurs ont été filmés avec en parallèle une vidéo mono-vue et un squelette 3D reconstruit à partir de capteurs de position placés sur le corps. Cela nous a permis d'obtenir un outil de reconstruction 3D efficace de la posture, avec comme résultat intermédiaire les données 2D issues d'OpenPose.

D'autres outils modernes [4] s'appuyant sur des réseaux de neurones convolutionnels nous ont permis d'obtenir directement une soixantaine de points du visage en 3D.

Le traitement des mains est quant à lui beaucoup plus délicat. En effet, les degrés de liberté sont nombreux, et le flou cinétique dû à la vitesse élevée des mains lors d'un discours en LSF ne simplifie pas les choses. Les outils les plus récents effectuent une classification basée sur l'apparence des mains, en déterminant la configuration manuelle la plus probable parmi un nombre fixé à l'avance. Les travaux d'Oscar Koller [5] notamment s'appuient sur des données de traduction de la météo en Allemagne.

Comme annoncé plus haut, le second volet de ce projet, à peine entamé, concerne le traitement automatique de la LSF à partir des informations extraites des vidéos. Une première approche vise à faire de la reconnaissance de signe isolé (en opposition à la détection de signe en contexte, qui est plus complexe). Des premiers résultats encourageants semblent montrer qu'avec une simple comparaison de signes en Dynamic Time Warping, la reconnaissance est possible. La poursuite de ces travaux a deux objectifs principaux : d'abord, permettre l'interrogation d'une base de données de LSF directement avec une vidéo de LSF, sans passer par une entrée textuelle ; ensuite, la reconnaissance de signe isolé peut être vue comme une étape pour réaliser de la détection de signes en contexte, d'éléments linguistiques, etc.

[1] Zhe Cao, Tomas Simon, Shih-En Wei et Yaser Sheikh : Realtime multi-person 2d pose estimation using part affinity fields. *In CVPR*, 2017.

[2] Ruiqi Zhao, Yan Wang, C Fabian Benitez-Quiroz, Yaojie Liu et Aleix M Martinez : Fast and precise face alignment and 3d shape reconstruction from a single 2d image. *In European Conference on Computer Vision*, pages 590–603. Springer, 2016.

[3] Limsi et CIAMS : Mocap1, 2017. URL `https://hdl.handle.net/11403/mocap1/v1`. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[4] Adrian Bulat et Georgios Tzimiropoulos : How far are we from solving the 2d & 3d face alignment problem ? (and a dataset of 230,000 3d facial landmarks). *In International Conference on Computer Vision*, 2017.

[5] Oscar Koller, Hermann Ney et Richard Bowden : Deep hand : How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3793–3802, juin 2016.