Utilisation de techniques de Deep Learning pour la reconnaissance de dystrophies neuromusculaires liées au FSHD à travers l'analyse des mouvements faciaux

Yoann Torrado & François Bremond

March 2025

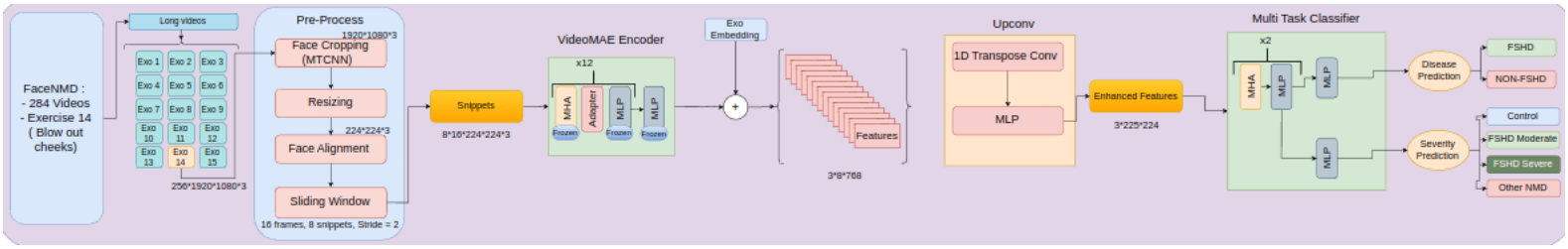

Figure 1: Notre architecture

Dans cette étude, nous explorons l'utilisation de techniques de deep learning dans le cadre de la reconnaissance des dystrophies neuromusculaires liées au FSHD. Notre méthode se focalise sur l'analyse des mouvements faciaux pour classifier la présence ou non du FSHD. Nous utilisons un jeu de données de 113 patients uniques récolté par le CHU de Nice. Le protocole de collecte permet de filmer 15 exercices faciaux avec deux caméras, ce qui nous donne 284 longues vidéos.

Classes	**Training Set**	**Testing Set**
FSHD Modéré	58	24
FSHD Sévère	30	14
Autre NMD	64	28
Contrôle	46	20

Table 1: Répartition des vidéos entre le training set et le testing set

Pour mener à bien notre étude, nous avons divisé les vidéos en un ensemble d'entraînement (training set) et un ensemble de test (testing set) en

utilisant une division stratifiée. Cette approche permet de conserver la même répartition des classes que dans l'ensemble de données initial, garantissant ainsi un échantillonnage équilibré.

Notre modèle est basé sur l'architecture de VideoMAE [1] et pré-entraîné sur la base de données VoxCeleb, permettant une initialisation solide des poids grâce à une large base de vidéos faciales. VideoMAE est utilisé en tant qu'extracteur de caractéristiques, et seul l'encodeur du transformer est utilisé lors de l'entraînement sur notre base de données.

Durant le pré-entraînement sur VoxCeleb, les clips sont partiellement masqués, puis les patches non masqués sont transmis à l'encodeur du transformer, qui extrait une matrice de descripteurs représentant les mouvements du visage au cours du clip. Cette matrice est ensuite passée au décodeur du transformer, chargé de reconstruire les clips. Cette tâche permet d'obtenir un encodeur capable d'extraire une matrice de descripteurs capturant efficacement les dynamiques faciales.

Avant le traitement de la base de données du CHU, un prétraitement est nécessaire. Les longues vidéos sont d'abord segmentées en 15 clips à l'aide du logiciel Elan, chacun lié à un exercice. Puis le visage est découpé (face cropping) afin d'éliminer l'arrière-plan pouvant gêner l'analyse, ensuite les clips sont redimensionnés à 224*224 pixels. Enfin, le visage est aligné à l'aide d'une technique de recalage d'image pour faciliter l'analyse spatiale du modèle. Un prétraitement temporel des clips est également nécessaire, car leur durée reste trop importante pour être directement exploitée par notre modèle. Pour cela, nous sélectionnons les 8 premiers snippets de chaque clip afin de nous concentrer sur le début du mouvement. Chaque snippet est composé de 16 images, échantillonnées à raison d'une image sur deux.

Lors de l'entraînement du modèle, les clips sont transmis à l'encodeur pré-entrainé du transformer, qui extrait une matrice de descripteurs représentant les mouvements du visage pendant le clip. Ensuite, la matrice est transmise à un module Upconv, composé d'une couche de déconvolution visant à augmenter sa résolution temporelle et d'un MLP permettant d'affiner les descripteurs. Cette étape améliore la matrice en optimisant la représentation temporelle des descripteurs. Finalement, cette matrice est classifiée à l'aide d'un transformer léger, composé de deux blocs d'attention et de deux têtes de classification. Ces deux têtes permettent de réaliser de l'apprentissage par multi-tâches: une tâche est binaire et permet de prédire si le clip est FSHD ou non et l'autre est une classification de la sévérité entre les 4 classes. De plus, l'utilisation d'adaptateur permet d'affiner l'encodeur tout en réduisant le coût computationnel du modèle.

Pour l'évaluation du modèle, nous avons choisi comme exercice représentatif l'exercice 14 : gonfler les joues. Nous avons comparé plusieurs configurations et observé une amélioration des performances avec l'utilisation des modules de Upconv, des adaptateurs et de l'apprentissage par multi-tâches. Le modèle basé sur l'extracteur de caractéristiques de VideoMAE a permis d'obtenir une précision de 79.1% pour la classification entre FSHD et non-FSHD. En cumulant les prédictions des 15 exercices avec un vote majoritaire, nous obtenons une précision de 72%.

Architectures	Pre-train Dataset	Précision (%)
I3D	Kinetics	48.84
VideoMAE	Kinetics	55.8
VideoMAE	VoxCeleb	74.4
VideoMAE + Upconv + Multi-task	VoxCeleb	76.7
OUR (VideoMAE + Adaptateur + Upconv + Multi-task)	VoxCeleb	79.1

Table 2: Comparaison des résultats des modèles de deep learning pour la détection de FSHD.

Lors de l'inférence, le modèle a pu traiter l'ensemble du testing set en moins d'une minute, offrant ainsi une utilisation rapide et efficace. L'analyse de la matrice de confusion révèle une bonne capacité à identifier les patients sains, avec une spécificité de 79,2%, bien que la sensibilité reste plus limitée à 63,2%.

L'étude démontre qu'il est possible, avec des modèles de deep learning, de détecter des dystrophies faciales dans des vidéos, mais les résultats restent très sensibles à la quantité et à la qualité des données. Malgré ces limitations, ces techniques ouvrent des perspectives pour des outils de diagnostic automatisés plus précis dans le suivi de patients atteints de FSHD.

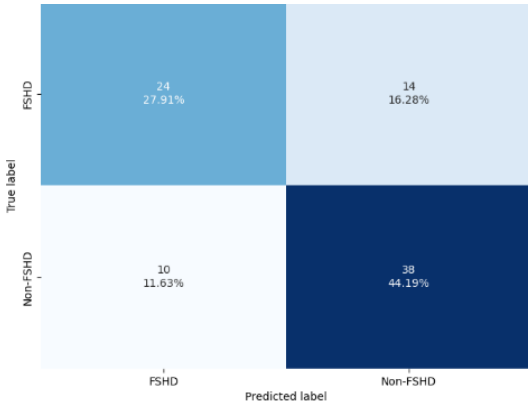

Figure 2: Confusion Matrice

References

[1] Jue Wang Zhan Tong, Yibing Song and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *In Advances in Neural Information Processing Systems.*, 2022.