

## Session 1 Comparaison de Classifieurs

L'objectif est de comparer deux approches très utilisées en vérification d'identité par la signature manuscrite en-ligne: la distance élastique ou « Dynamic Time Warping » (DTW), et une approche basée sur la modélisation statistique par Modèles à Mélanges de Gaussiennes ou « Gaussian Mixture Models » (GMMs).

### I. La base de signatures

#### 1. Description de MCYT-100

La base de signatures MCYT contient les signatures de 100 personnes labélisées de 0 à 99. Elles ont été acquises sur une tablette à digitaliser WACOM Intuos2 A6, avec un stylo spécial à encre, à une fréquence d'échantillonnage de 100Hz. En chaque point d'une signature, la tablette capture les coordonnées  $(x, y)$ , la pression du stylo  $p$  et deux angles (azimuth et altitude), angles qui encodent la position du stylo dans l'espace. Pour chaque personne  $i$ , 25 signatures authentiques (étiquetées  $ivXX$ ) et 25 signatures imitées dites « skilled forgeries » (étiquetées  $ifXX$ ) ont été acquises.

Chaque signature est représentée comme une matrice dans un fichier (.txt), où chaque ligne correspond à un point de sa trajectoire, comme suit:

Coordonnée X   Coordonnée Y   Pression   Altitude   Azimuth

#### 2. Visualisation des signatures d'une personne

Le programme Matlab "**visualize.m**" permet de visualiser toutes les signatures (authentiques et imitées) d'une personne. Sur Matlab, utilisez la commande **visualize(sig)** pour voir les signatures de la personne **sig**, où **sig** est le numéro de l'utilisateur (de 0 à 99).

### II. Le classifieur à distance élastique ("Dynamic Time Warping")

Dans cette partie, nous allons considérer seulement 8 personnes: celles numérotées 11,15,16,17,18,20,21,22. Pour chacune, on dispose de 25 signatures authentiques et 25 imitations. A ce stade, nous travaillerons seulement sur la description des signatures par leurs coordonnées  $(x,y)$ .

Exécutez le script "**Main\_DTW.m**" (qui appelle la fonction "**DTWtest.m**") réalisant pour chaque personne, l'appariement ("matching") ou calcul de la distance élastique entre sa **première signature** (prise comme **signature de référence**) à ses autres 24 signatures authentiques (prises alors comme **signatures de test**) et aux 25 signatures imitées (prises comme impostures de **test**). Ce script affiche la distribution des distances DTW: **en bleu, les scores des 24 authentiques** et **en rouge les 25 scores imposteur**.

1. Comparer ces distributions obtenues sur les 8 personnes en termes du **recouvrement** entre les scores CLIENT et les scores IMPOSTEUR. Est-il similaire pour toutes ces 8 personnes? Analysez chaque cas (personne) après avoir visualisé ses signatures. Qu'observez-vous?
2. Même question en utilisant la description des signatures par  $(x,y,p)$  et  $(x,y,p,Az,Alt)$ , et comparer le résultat à ceux que vous avez obtenus en question 1. Qu'observez-vous ?
3. Exécutez le script **Intra\_class\_variability.m** qui affiche la variabilité intra-classe des 5 premières signatures authentiques de chaque personne avec la description  $(x,y)$  (en considérant les mêmes 8 personnes). Quelles sont les personnes ayant les signatures les plus stables?

4. Avec le même script, affichez la variabilité intra-classe des signatures authentiques de chaque personne avec la description  $(x,y,p)$  et  $(x,y,p,Az,Alt)$ . Est-ce que cela a un impact sur la stabilité des signatures?

### III. Le classifieur à Mélange de Gaussiennes (« Gaussian Mixture Model »)

Dans cette partie, nous allons considérer les mêmes 8 personnes: celles numérotées 11,15,16,17,18,20,21,22. Pour chacune, on dispose de 25 signatures authentiques et 25 imitations.

Exécutez le script “**main\_GMM.m**” pour entraîner pour chaque personne  $i$  son modèle sur ses 5 premières signatures authentiques, numérotées de 0 à 4. Le nombre de densités Gaussiennes est calculé comme le nombre total de points dans les 5 signatures toutes confondues, divisé par 100 (nombre de points utilisé pour estimer les paramètres de chaque composante gaussienne).

La Log-Vraisemblance est calculée sur les 20 signatures authentiques restantes (n'appartenant pas à l'ensemble d'apprentissage) et sur 20 signatures imitées ou « skilled forgeries ». Analysez les histogrammes des scores de Vraisemblance des 20 signatures authentiques (affichées en bleu) et des 20 imitations (affichées en rouge) sur la même figure.

1. Comparer par personne la qualité du modèle en considérant  $(x,y)$ ,  $(x,y,p)$ ,  $(x,y,p,Az,Alt)$ .
2. Est-ce que le GMM est un bon modèle? Comment pourriez-vous l'améliorer?