

École de Technologie Supérieure  
SYS800: Reconnaissance de formes et inspection  
Laboratoire 1: Représentation, extraction et sélection de  
caractéristiques

M'Hand Kedjar  
Code ÉTS: KEDM09088002

21 octobre 2016

# I Introduction

Le but de ce premier laboratoire est l'extraction de caractéristiques à partir d'une base de données composée des chiffres manuscrits (de 0 à 9).

Nous disposons de 7000 chiffres manuscrits répartis en 10 classes représentant les chiffres 0 jusqu'à 9, à partir desquels on extrait 2 sous-ensembles. Le premier contient 600 exemples de chacun des chiffres, qui seront utilisés dans la phase d'entraînement, et le second contient 100 exemples de chaque classe qui seront utilisés dans la phase de test.

## II Prétraitement et préparation des données : extraction des caractéristiques

### II.1 Codage de la fonction *zoneProjection.m*

Nous allons utiliser le logiciel Matlab pour programmer notre solution. La première étape consiste à coder la fonction *zoneProjection.m* qui implémente la méthode de projection de zones est qui est composée des sous-étapes suivantes :

- Redimensionner l'image de départ en taille  $T_{img} = m \times n$
- Diviser l'image redimensionnée en plusieurs zones  $T_Z = z_r \times z_c$  chacune de taille  $T_{zone} = m_z \times n_z$
- Pour chaque zone dénotée  $Z_k, k = 1, \dots, T_Z$ , nous la projetons verticalement pour former son vecteur représentatif  $d$  de  $n_z$  composantes. Chaque composante de  $d$  est la somme de la valeur des pixels de la colonne correspondante :  $d_j = \sum_{i=1}^{m_z} Z_{kij}, j = 1, \dots, n_z$
- La valeur de  $Z_k, k = 1, \dots, T_{zone}$ , est égale à la valeur moyenne des composantes de chaque vecteur  $d$  :  $Z_k = \frac{1}{n_z} \sum_{j=1}^{n_z} d_j$ . Nous obtenons donc une matrice  $Z$  de taille  $T_Z = z_r \times z_c$ , que nous allons transformer en un vecteur  $v_Z$  en concaténant les colonnes de  $Z$  :  $v_Z = [Z_1, \dots, Z_{z_r}]^T, \dots, [Z_{z_r \times (z_c-1)+1}, \dots, Z_{z_r \times z_c}]^T$
- Nous calculons un deuxième vecteur  $v_{z_r}$  qui représente la valeur moyenne de chaque ligne de  $Z$ , et un troisième vecteur  $v_{z_c}$  qui représente la valeur moyenne de chaque colonne de  $Z$ .
- Le vecteur caractéristique final de l'image sera formé en concaténant les 3 vecteurs ainsi obtenus :  $v_{image} = [v_Z \quad v_{z_r} \quad v_{z_c}]$

Ces étapes sont illustrées en figure 1.

### II.2 Variation des paramètres de la méthode

Dans cette section, nous allons générer diverses bases de données en faisant varier les paramètres de la méthode : taille de l'image :  $N_p = m \times n$ , et taille de la zone  $T_{zone} = m_z \times n_z$ .

Une des mesures les plus importantes pour évaluer la capacité de discrimination d'une méthode d'extraction de caractéristiques est le taux de recouvrement  $r$ . Ce taux est calculé de la façon suivante :

- Pour chaque vecteur caractéristique  $X_k, k = 1, \dots, n$ , nous calculons sa distance Euclidienne avec tous les autres vecteurs caractéristiques.
- Nous notons le vecteur  $X_u$  celui qui a la plus petite distance Euclidienne avec  $X_k$ .
- Si  $X_u$  n'est pas de la même classe que  $X_k$ , il y a recouvrement :  $r_k = 1$ , sinon  $r_k = 0$ .
- le recouvrement final est donné par  $r = \frac{1}{n} \sum_{k=1}^n r_k$

Le tableau 1 montre la valeur du taux de recouvrement  $r$  en fonction des paramètres de la méthode :  $T_{img}$  : taille de l'image,  $T_{zone}$  : taille de la zone,  $r$  : taux de recouvrement,  $N_p$  : nombre de pixels de l'image, et  $N_c$  : nombre de composantes de chaque vecteur caractéristique.

Nous remarquons que le meilleur taux de recouvrement est égal à 1.72% et cette valeur est obtenue pour les

$T_{img}$	$T_{zone}$	$r$	$N_p$	$N_c$
100 × 80	10 × 10	<b>1.72%</b>	8000	98
	5 × 5	1.85%	8000	356
60 × 50	10 × 10	2.12%	3000	41
	5 × 5	1.75%	3000	142

TABLE 1 – Évolution du taux de recouvrement en fonction des paramètres de la méthode.

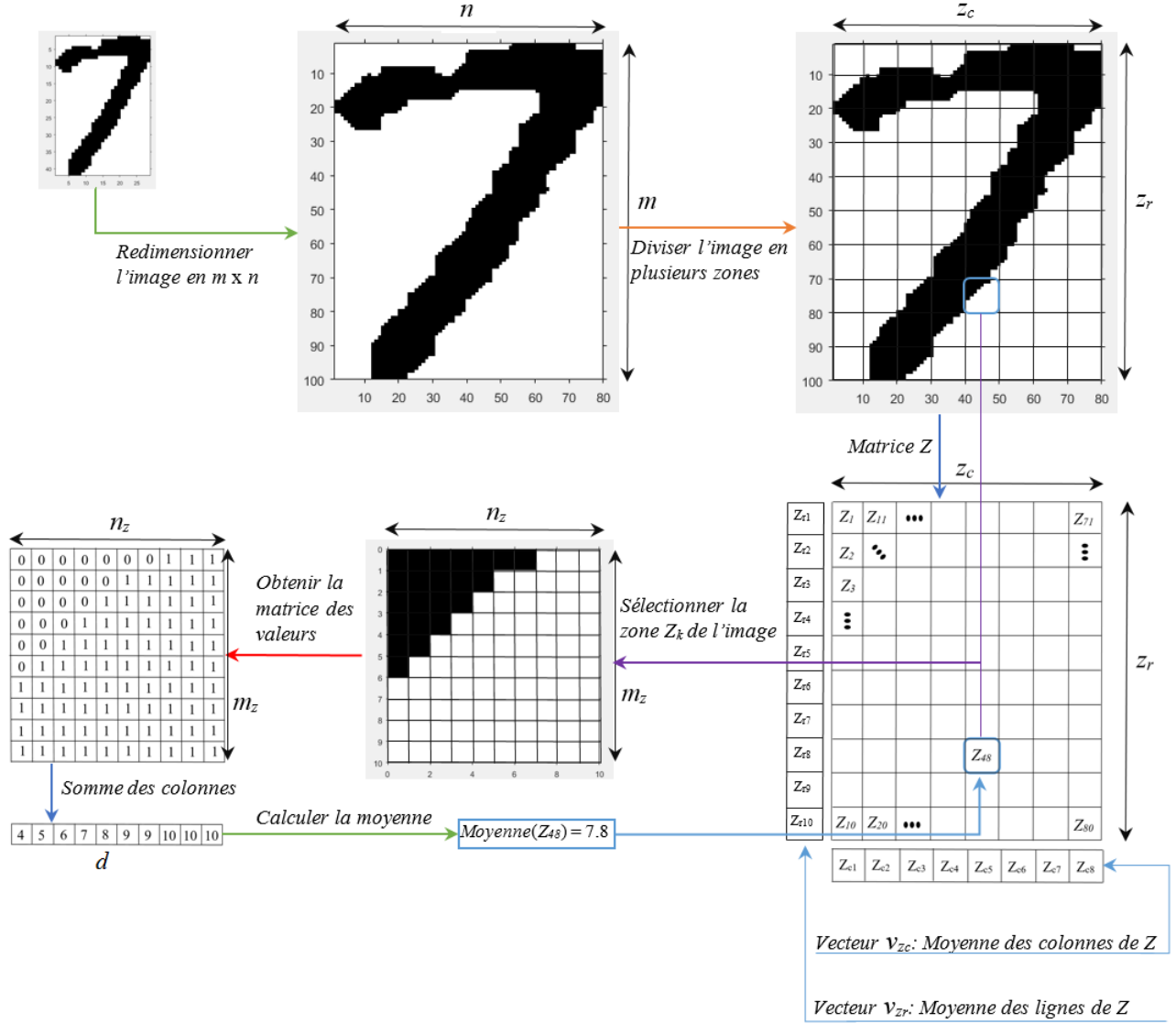


FIGURE 1 – Extraction des caractéristiques

paramètres suivants ;  $m = 100, n = 100, m_z = 10, n_z = 10$ .

Nous remarquons que le taux de recouvrement dépend fortement de la taille de l'image et du nombre de composantes des vecteurs caractéristiques. Le taux de recouvrement augmente dans les deux cas suivants :

- Si la taille de l'image est petite et qu'on a des grandes zones : On ne capture pas assez d'information pour pouvoir discriminer entre les classes.
- Si la taille de l'image est grande et qu'on a des petites zones : On capture beaucoup trop d'information, et l'effet du "bruit" peut se ressentir.

Ainsi, pour avoir le meilleur taux de recouvrement, on a besoin de résoudre un problème d'optimisation à plusieurs paramètres.

Pour comprendre quelles sont les classes qui se recouvrent le plus, nous avons analysé le taux de recouvrement de chaque chiffre (0, 1, ..., 9) par rapport aux autres chiffres, et ce pour les quatre configurations précédentes. La figure 2 montre le résultat obtenu.

En Analysant plus en détail le résultat de la figure 2, nous confirmons que le taux de recouvrement est élevé entre les chiffres (3 et 8), (7 et 9), (9 et 8), pour ne citer que les plus importants. De plus, nous remarquons que dans la configuration 3 ( $m = 60, n = 50, m_z = 10, n_z = 10$ ), il y a un grand recouvrement entre le chiffre 9 et le chiffre 7.

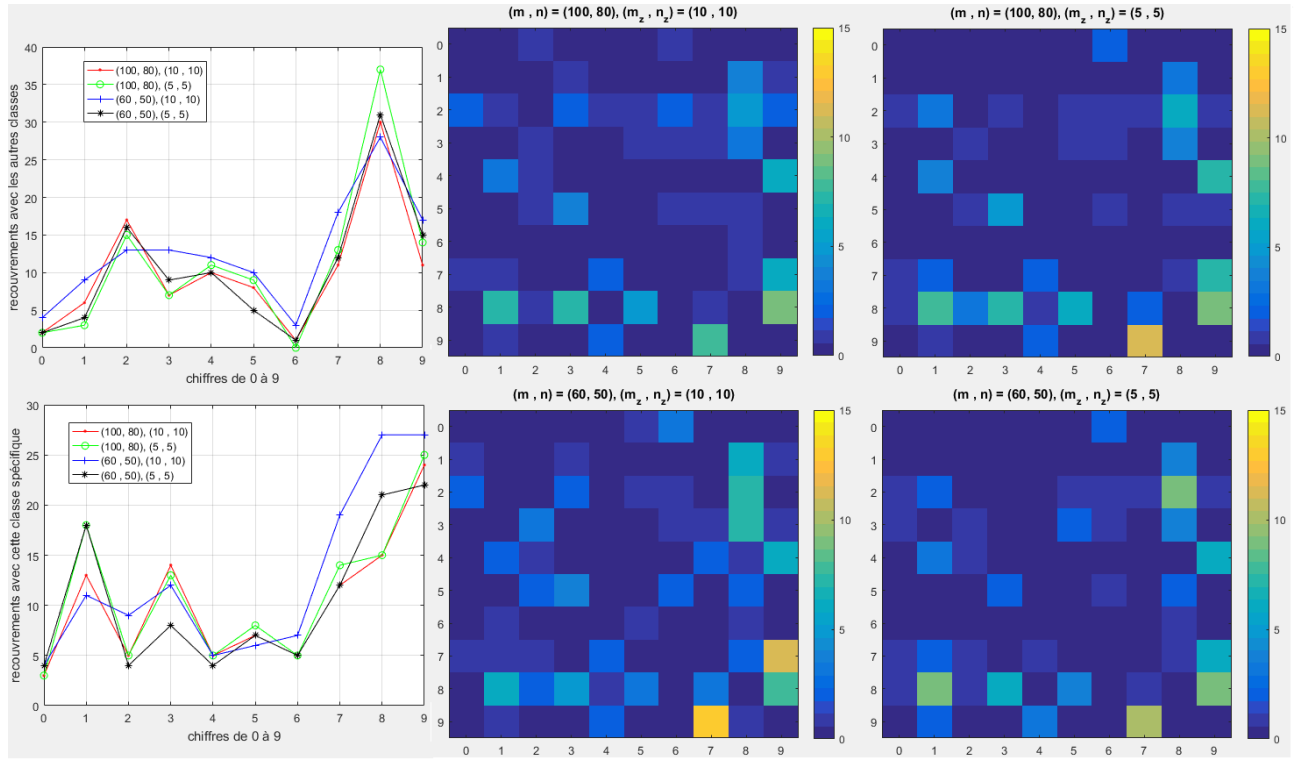


FIGURE 2 – Comparaison du taux de recouvrement

Dans la figure 2, les courbes en haut à gauche représentent le recouvrement d'un chiffre donné avec les toutes les autres classes. Par exemple, le chiffre 8 dans la configuration  $(m = 100, n = 80, m_z = 5, n_z = 5)$  a un recouvrement de  $\frac{37}{6000}$  avec les autres classes.

Les courbes en bas à gauche représentent le recouvrement des autres chiffres avec une classe spécifique. Gardons toujours le même exemple, c'est-à-dire le chiffre 8 dans la configuration  $(m = 100, n = 80, m_z = 5, n_z = 5)$ , les autres chiffres ont un recouvrement de  $\frac{15}{6000}$  avec le chiffre 8.

Ces courbes confirment que le taux de recouvrement est plus élevé pour les chiffres 7, 8 et 9, et ce spécialement pour la configuration 3 qui a le taux de recouvrement le plus élevé  $r = 2.12\%$ .

### II.3 Extraction des caractéristiques de la base de test

Nous avons démontré que la configuration 1 ( $m = 100, n = 80, m_z = 10, n_z = 10$ ) donne le meilleur taux de recouvrement ( $r = 1.72\%$ ). Nous allons donc utiliser ces paramètres pour extraire les caractéristiques de la base de test. Après avoir extrait les caractéristiques, nous avons calculé le taux de recouvrement correspondant. Nous avons trouvé la valeur  $r = 2.9\%$ . Cette valeur est plus élevée que celle trouvée avec la base d'apprentissage ( $r = 1.72\%$ ). Ceci peut s'expliquer par le fait qu'il y a moins de données dans la base de test : 100 exemples dans chaque classe comparés aux 600 pour la base d'entraînement. Cependant, il est important de noter que le taux de recouvrement obtenu avec la base de test est tout à fait raisonnable.

## III Réduction de la dimension par l'ACP

### III.1 Réduction de la dimension de l'espace des caractéristiques

Dans cette section, nous allons utiliser l'Analyse en Composantes principales (ACP) afin de réduire la dimension de l'espace des caractéristiques.

À partir de l'étape de l'extraction des caractéristiques, nous avons obtenu une matrice  $B$  de 6000 lignes et 98

colonnes.

Les étapes pour calculer l'ACP avec Matlab sont les suivantes :

- Calculer la moyenne de la matrice  $B$  :  $M = \text{mean}(B)$
  - Calculer la matrice de covariance de  $B$  :  $C = \text{cov}(B)$
  - Calculer les vecteurs et les valeurs propres de  $C$  :  $[v, \lambda] = \text{eig}(C)$
  - Trier les vecteurs propres selon les valeurs propres (en ordre croissant)
  - Retenir les  $k$  premiers vecteurs propres qui respectent une certaine condition de seuil  $s$  tel que  $\frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^d \lambda_j} \geq s$ , avec  $d$  le nombre de vecteurs dans l'espace de départ.
  - Projeter sur les axes principaux correspondant aux  $k$  vecteurs propres retenus :  $(B - M) * v_k$
- Nous avons décidé de prendre comme seuil  $s = 95\%$ . La valeur de  $k$  correspondante est  $k = 44$

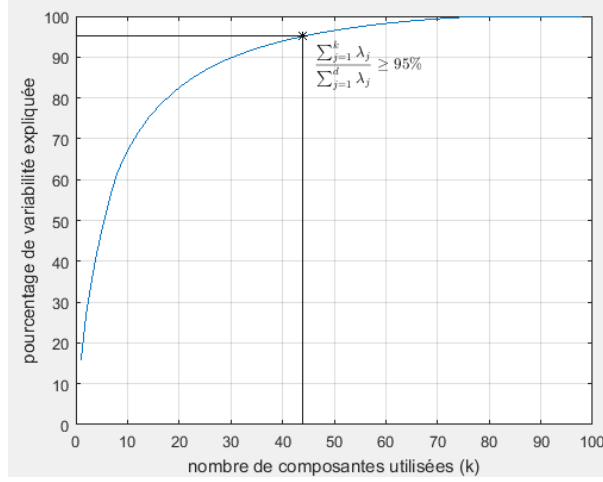


FIGURE 3 – Détermination du nombre de composantes à retenir

La figure 3 donne le pourcentage de la variabilité expliquée en fonction du nombre de composantes utilisées ( $k$ ). Le point ( $k = 44, 95.11\%$ ) est illustré. On trouve donc qu'utiliser 44 vecteurs propres des 98 originaux permet de garder une variabilité de données de plus de 95%.

Nous avons aussi calculé le taux de recouvrement dans cette nouvelle base. En utilisant la projection des vecteurs caractéristiques sur les 44 premiers vecteurs propres obtenus, on trouve un taux de recouvrement de  $r_{acp} = 1.62\%$ . Cette valeur est meilleure que celle trouvée auparavant ( $r = 1.72\%$ ), ce qui est compréhensible vu que le but de l'ACP est de chercher une projection pour minimiser au maximum le recouvrement. De même, si nous calculons le taux de recouvrement de la base de test après projection, on trouve  $r_{testacp} = 2.80\%$  qui est inférieur par rapport au taux trouvé précédemment ( $r = 2.9\%$ ). Pour comprendre en détail l'influence du nombre de composantes principales sur le taux de recouvrement, nous avons tracé dans la figure 4  $r$  en fonction de  $k$ , et ce pour les bases d'apprentissage et de test.

Nous remarquons que pour des petites valeurs de  $k$ ,  $r$  est assez élevé. Pour la base d'apprentissage,  $r$  atteint la valeur optimale (1.53%) pour  $k = 37$ , alors que la base de test la valeur optimale (2.5%) est atteinte pour  $k = 18$ . Ce que nous pouvons conclure est que les valeurs optimales pour les bases d'apprentissage et la base de test ne sont pas atteintes pour la même valeur de  $k$ . De plus, à partir d'une certaine valeur de  $k$ , le taux de recouvrement pour la base d'apprentissage est toujours plus bas que celui de la base de test. Si nous rajoutons la contrainte du seuil  $s \geq 95\%$  qui est atteinte pour  $k = 44$ , nous pouvons conclure que le choix de  $k$  dépend de plusieurs facteurs, entre autres le taux de recouvrement souhaité pour les bases d'apprentissage et de test, le pourcentage de variance des données que nous devons capturer, et aussi le temps de calcul et les ressources matérielles dont nous disposons.

Le tableau 2 résume les valeurs importantes de  $k$  pour ce qui est du taux de recouvrement ( $r_a$  : taux pour l'ensemble d'apprentissage,  $r_t$  : taux pour l'ensemble de test) et du seuil  $s$  choisi. Il est important de remarquer que si on utilise la totalité des vecteurs propres ( $k = 98$ ) pour la projection, on retrouve le taux de recouvrement original ( $r = 1.72\%$ )

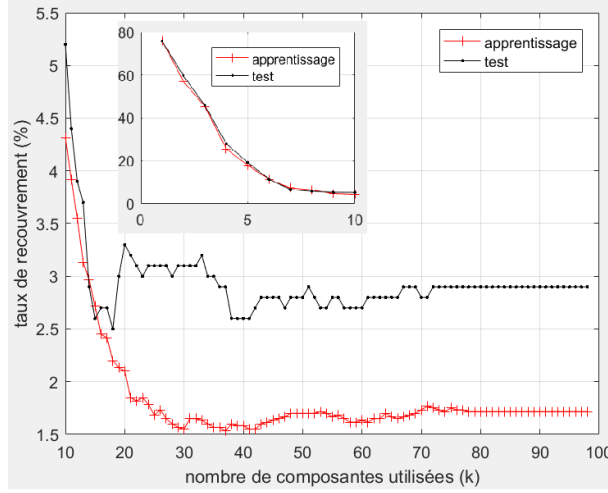


FIGURE 4 – Détermination du nombre de composantes à retenir

k	$r_t$	$r_a$	commentaires
18	2.5%	2.2%	Meilleur taux pour le test
37	2.9%	<b>1.53%</b>	Meilleur taux pour l'apprentissage
44	2.8%	1.62%	Taux pour le seuil $s \geq 95\%$
98	2.9%	1.72%	Taux en prenant la totalité des vecteurs propres

TABLE 2 – Taux de recouvrement en fonction de  $k$

### III.2 Extraction des caractéristiques des images de la base de test

Nous avons donc tous les paramètres pour extraire les caractéristiques des images de la base de test. À partir de la fonction `zoneProjection.m`, nous avons  $m = 100, n = 80, m_z = 10, n_z = 10$ , et nous avons obtenu une matrice  $T$  de 1000 lignes et 98 colonnes. À partir du calcul des vecteurs propres des données d'apprentissage, nous avons  $k = 44$ . Pour extraire les caractéristiques de la base de test, nous allons projeter sur le nouvel espace formé par les  $k$  premiers vecteurs propres sélectionnés. Les étapes sont comme suit :

- Calculer la moyenne de la matrice  $T$  colonne par colonne :  $M = \text{mean}(T)$
- Soustraire  $M$  de  $T$  :  $T - M$
- Multiplier par les  $k$  premiers vecteurs propres :  $(T - M) * v_k$

Après la projection, nous allons avoir une nouvelle matrice de taille 1000 lignes et 44 colonnes. À ce stade, nous avons extrait les caractéristiques des deux bases de données qui vont nous servir dans l'étape de classification.

## IV Conclusion

Dans ce premier laboratoire, nous avons implémenté une solution pour extraire les vecteurs caractéristiques à partir d'une base de données de chiffres manuscrits. Nous avons dans un premier temps extrait les caractéristiques de la base de données d'apprentissage en faisant varier les paramètres de la méthode pour obtenir le meilleur taux de recouvrement. Nous avons ensuite utilisé ces mêmes paramètres optimaux pour extraire les caractéristiques de la base de données de test. Par la suite, nous avons utilisé l'Analyse en Composantes Principales (ACP) pour réduire la dimension de l'espace des caractéristiques. En ayant un seuil de 95%, nous avons trouvé qu'il faut prendre au moins  $k = 44$  vecteurs propres sur les 98. Nous avons aussi remarqué que le taux de recouvrement dépend de plusieurs facteurs : la taille de l'image, la taille de la zone et le nombre de composantes principales prises en compte. Ce laboratoire nous a démontré que le prétraitement et de la compréhension des données est d'une grande importance dans tout problème de classification. C'est une étape essentielle.