# RAPPORT DU PROJET 3
# Études et expérimentation de la classification des scènes naturelles

Rédigé par :
Nguyen Van Tho & Nguyen Quoc Khai

Sous la supervision de :
Mr HO Tuong Vinh

6 janvier 2014

## 1 Introduction

La reconnaissance et la catégorie des images sont importantes pour accéder à l'information visuelle au niveau d'objets et de types de scènes. Pour l'instant, un des types de reconnaissances le plus difficile est la reconnaissance des scènes naturelles. En fait, ce type de reconnaissance est défié par les défis tel ques la variation de point de vue, la différence d'échelle, le changement d'illumination, le désordre du fond, l'occlusion et la déformation.

La reconnaissance des scènes naturelles ont beaucoup d'application dans plusieurs domaines notamment la recherche d'images, la reconnaissance de location, ...

## 2 État de l'art

Plusieurs techniques ont été proposées au cours de ces dernières années afin de faciliter la classification des scènes naturelles. Les méthodes au dessous permet décrire les idées principales des méthodes étudiées au cours la durée de ce TP.

## 2.1 Approche par mots visuels « Visual Word »

### Caractéristiques :

On peut voir l'idée de cette méthode par l'image ci-dessous :

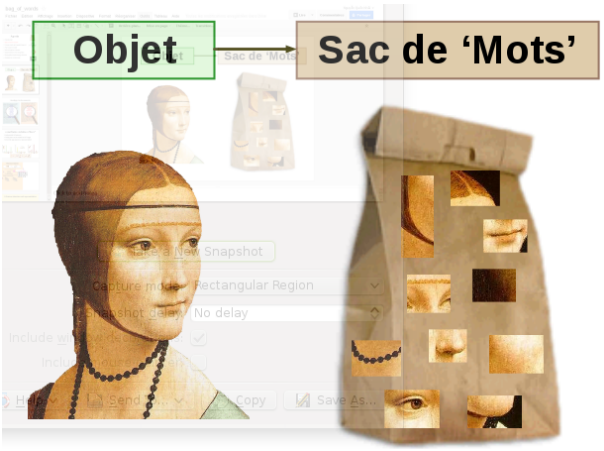


FIGURE 1 – Idée de la méthode « mots visuels »

Cette méthode obtient des caractéristiques locales de l'image : région « stables », points d'intérêt (SIFT, SURF...), etc. Ensuite, elle trouve des caractéristiques qui se répètent dans les images (construction d'un vocabulaire). Dans cette méthode, une image est décrite par les mots visuels qu'elle contient. Ces mots ne sont pas dans l'ordre. Dans cette méthode, on considère une image comme un sac de mots visuels.

### Algorithme

On peut décrire la méthode mots visuels en 3 étapes :

1. Trouver les caractéristiques stables des images
2. Construire un dictionnaire par calculer les histogrammes
3. Décrire chaque image comme un sac de mots visuels

### Reconnaissance par mots visuels

Après avoir détecté des caractéristiques et construit d'un dictionnaire, cette méthode peut reconnaitre ou classifier des objets en comparant le sac de mots de chaque image de test avec les sacs de mots des images d'apprentissage.

La méthode mots visuels peut être résumé par cette image :

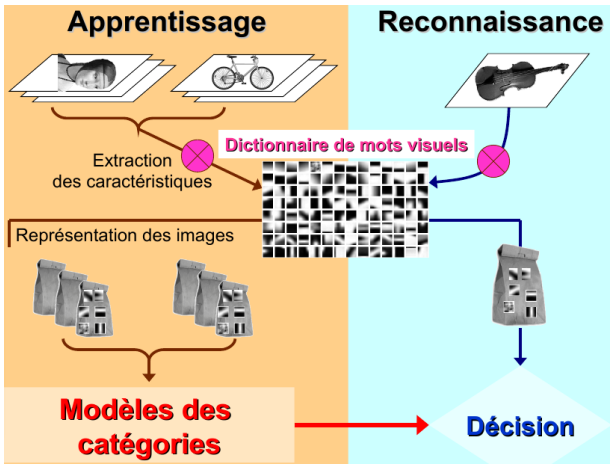


FIGURE 2 – Résumé de la méthode « mots visuels »

## 2.2 Méthode Beyond Bags of Features : Spatial Pyramid Matching

Basée sur la même idée avec la méthode bag of visual word, cette méthode prend aussi des points intérêts et construit les vocabulaires. Autrement dit, la méthode «Spatial Pyramid Matching» est différente dans l'étape de faire la correspondance.

### Pyramid Match Kernels

Cette méthode divise l'image en des bloques pour faire la correspondance et en des niveaux différents. Dans les niveaux différents, le nombre de bloques divisés est différent. Le nombre de bloques est augmenté quand le niveau augmente. $l$ signifie le niveau. $l = 0, 1, 2, ..., L$. Une fois que l'on a calculé les histogrammes, le nombre de correspondance est calculé par l'histogramme d'intersection :

$$I(H_X^l, H_Y^l) = \sum_{i=1}^{D} min(H_X^l(i), H_Y^l(i)) \tag{1}$$

*X, Y sont deux ensembles de points que l'on veut faire la correspondance.*
*$H_X^l, H_Y^l$ sont les histogrammes (nombre de points) de X et de Y au niveau $l$.*
*$D$ est le nombre de bloques et $H_X^l(i)$ est l'histogramme de $X$ au niveau $l$ dans le bloque $i$*
*Pour plus simple, on écrit $I^l$ au lieu de $I(H_X^l, H_Y^l)$*

Pour calculer le $I$ au niveau $L$, on applique exactement cette formule, mais pour les autres niveaux $l = 0, 1, ..., L-1$, on recalcule le nouveau $I$ comme ci-dessous :

$$I^l = I^l - I^{l+1} \tag{2}$$

Après avoir calculé tous les $LI$ pour tous les niveaux, on calcule $I$ total comme cette formule :

$$I = \sum_{l=0}^{L} \frac{1}{2^{L-l}} I^l \tag{3}$$

On peut voir un exemple de calculer les histogrammes des trois niveaux dans l'image ci-dessous :

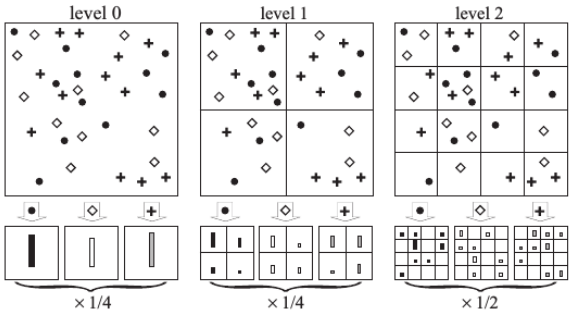


FIGURE 3 – Exemple de trois niveaux de la méthode Pyramid Matching

# 3 Méthodes proposées

En cherchant une méthode qui rend un résultat acceptable et qui est faisable dans le carde un projet universitaire, nous avons choisi parmi les méthodes présentées dans la partie l'état de l'art la méthode "bag of visual word". En effet, cette méthode est une combinaison de plusieurs méthodes. La méthode a des étapes suivantes :

- Détection des points d'intérêt et ses descripteurs
- Regroupement (Cluster) ces descripteurs pour définir un dictionnaire des mots, chaque centroïde est un mot
- Calcul les histogrammes de mots des images d'apprentissage selon le dictionnaire
- Classification les images de test en utilisant les histogrammes de mots

À chaque étape, on peut utiliser des méthodes différente. Nous choisissons la méthode SIFT afin d'extraire les points d'intérêt et ses descripteurs. En suite, la méthode Kmeans [1] est utilisée pour le regroupement des descripteurs afin de définir un dictionnaire. Pour la tâche de classification nous choisissons la méthode support vector machine (SVM) [2] qui rend un bon résultat et qui est vérifié dans l'était de l'art. Nous avons expérimenté 2 noyaux : linaire et RBF (Gaussien), le résultat de notre expérimentation montre que le noyau gaussien est plus approprié au ce genre de problème. Donc, nous ne présentons que le résultat avec le noyau gaussien.

# 4 Expérimentation

Nous avons fait plusieurs expérimentation afin d'évaluer notre approche et d'examiner l'impact du nombre de scène et de nombre des images sur le résultat. D'abord, nous fixons le nombre de mot du vocabulaire de 1000 et trouver les valeur optimale des autres paramètres de SVM. En suite, en utilisant ces valeurs optimales, nous expérimentons plusieurs valeur de taille du vocabulaire.

## 4.1 Critères d'évaluation

Dans le but d'avoir une vue plus précise des performances de notre programme, nous utilisons les indicateurs suivants :

- Matrice de confusion, le taux de rappel et le taux de précision de chaque scène. Cela est pour une analyse plus détaille
- Taux de reconnaissance (taux) qui est le rapport du nombre de bonnes reconnaissances par le nombre total.

– Temps d'apprentissage du programme
– Temps de reconnaissance

## 4.2 Méthode d'évaluation

Nous appliquons la méthode d'évaluation k-fold cross validation afin d'assurer que notre programme va bien prédire les nouvelles données. Nous choisissons 3 pour la valeur de k. Autrement dit, nous divisons les données en trois ensembles et chaque expérimentation nous prenons un ensemble la validation et deux autres ensembles pour l'apprentissage. En effet, avant de diviser les données, nous avons les ordonné aléatoirement afin de recevoir un résultat plus précis. La figure 4 l'explique notre méthode d'évaluation. Le résultat présenté dans ce rapport est la moyenne de 3 fois d'expérimentations.

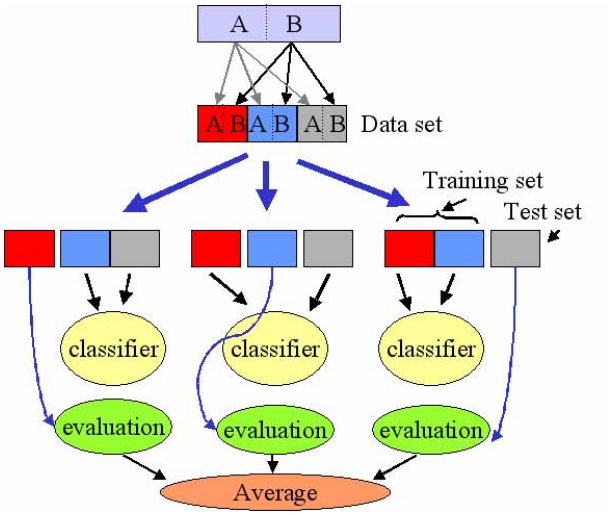


FIGURE 4 – Méthode 3-fold cross validation utilisée

## 4.3 Scénarios

### Expérimentation avec une partie des données

Vu que le temps de calcul de programme est élevé, nous l'expérimentons pour but de trouver les valeurs optimales de c et gamma de SVM. De plus, nous souhaitons examiner l'impact de nombre de scène en constatant que le taux de reconnaissance diminue quand le nombre de scène augmente. Le tableau ci-dessous présente les paramètres que nous utilisons pour cette expérimentation. D'abord, nous expérimentons avec la valeur fixée de C (300) alors que la valeur de gamma est variée de 0.05 à 2.0. La valeur optimale de gamma est ensuite utilisée pour trouver la valeur optimale de C.

| C | gamma | Nombre de scène | Nombre d'image / scène |
|---|---|---|---|
| 300 | 0.05 0.1 0.5 1.0 2.0 | 13 | 120 |
| 10 100 200 300 400 500 | gamma optimale | 13 | 120 |

Au vu des les taux de reconnaissance dans la figure ci-dessus, nous pouvons choisir la valeur optimale pour gamma. Avec la une valeur de gamma de 0.5, nous obtenons le taux maximum de

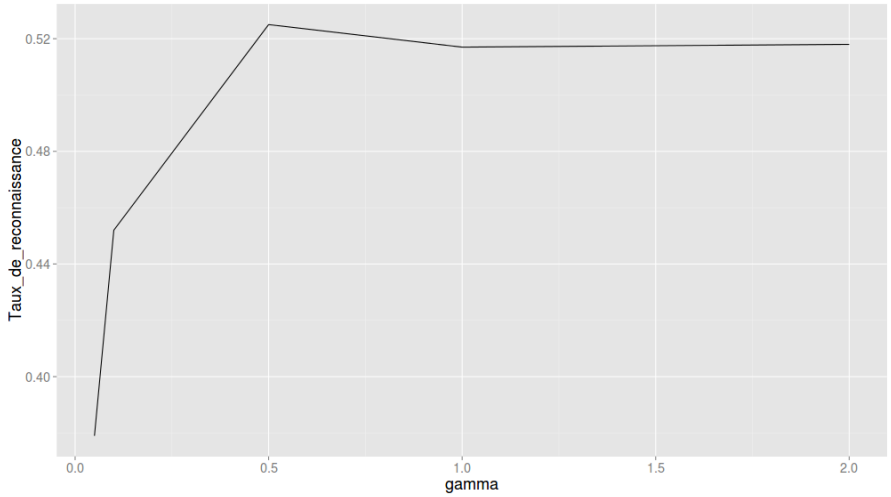


FIGURE 5 – Taux de reconnaissance avec une valeur fixée de C = 300 et des variation de gamma

reconnaissance. Nous allons utiliser cette valeur pour l'étape suivante : trouver la valeur optimale de C.

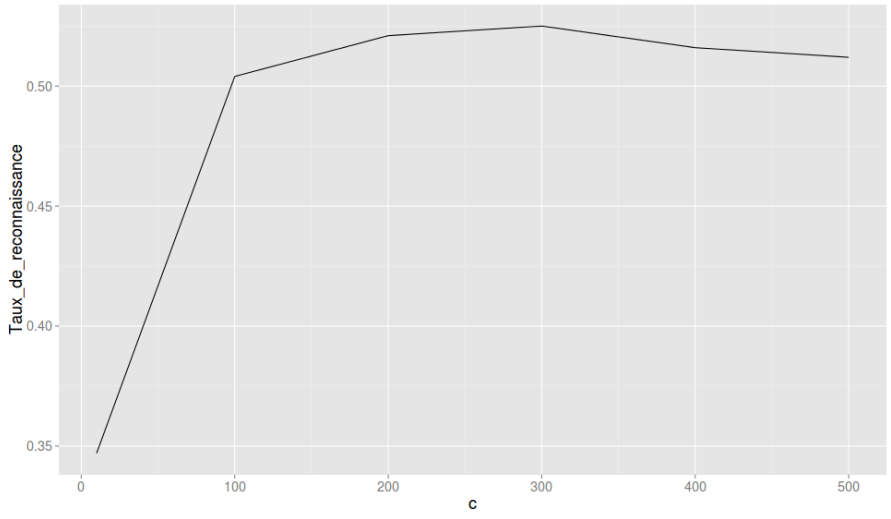


FIGURE 6 – Taux de reconnaissance avec une valeur fixée de gamma = 0.5 et des variation de C

En voyant la figure ci-dessus nous constatons que la valeur optimale de C est 300. Nous allons utiliser cette valeur de C et la valeur de gamma = 0.5 pour les autres expérimentations.

### Expérimentation avec les valeurs différentes de la taille de vocabulaire

La taille de vocabulaire influence beaucoup le taux de reconnaissance [3]. Nous voulons donc expérimenter plusieurs valeurs de la taille de vocabulaire afin de trouver la meilleure valeur. De plus, nous expérimentons le programme avec toutes les données et avec les paramètres optimales.

| Taille de vocabulaire | C | gamma | Nombre de scène | Nombre d'image / scène |
|---|---|---|---|---|
| 500, 1000, 2000, 5000 | 300 | 0.5 | 13 | maximum |

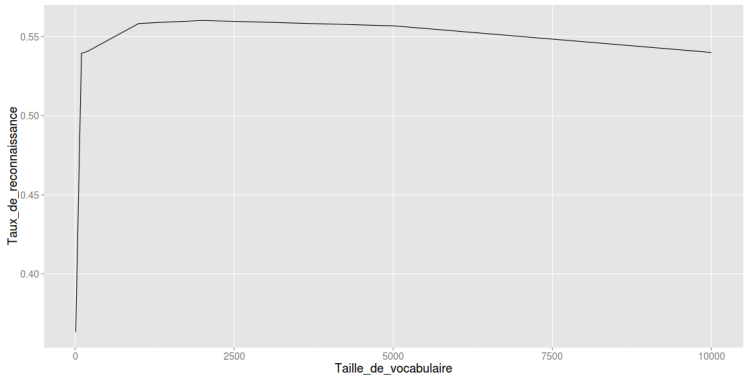


FIGURE 7 – Taux de reconnaissance avec des tailles différentes du vocabulaire

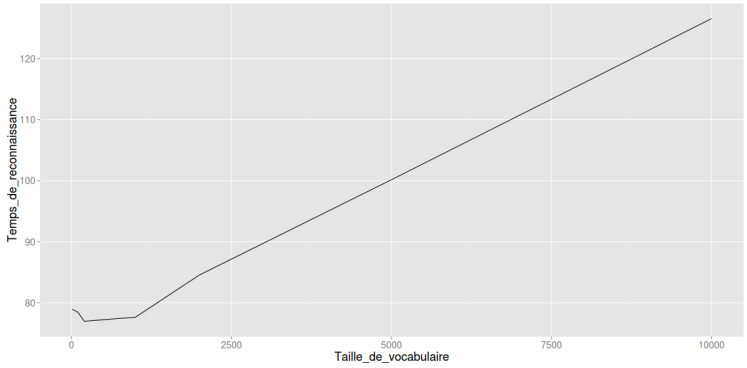


FIGURE 8 – Temps de reconnaissance avec des tailles différentes du vocabulaire

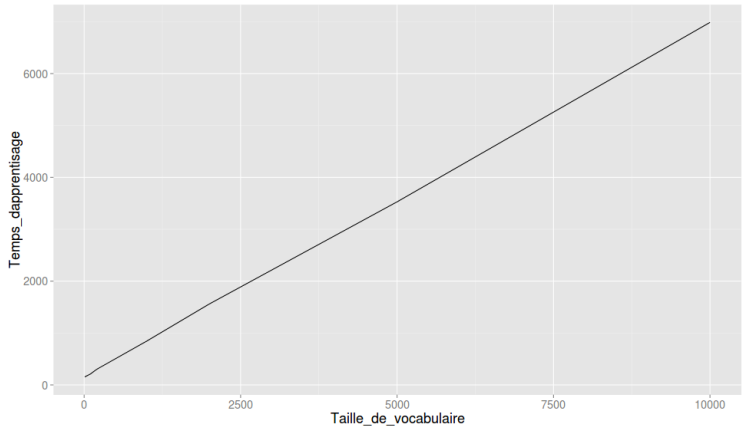


FIGURE 9 – Temps d'aprentissage avec des tailles différentes du vocabulaire

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | Taux de rappel(%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 69 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 2 | 1 | 0 | 0 | 6 | 85.2 |
| 2 | 8 | 47 | 5 | 13 | 2 | 9 | 20 | 4 | 2 | 5 | 2 | 0 | 3 | 39.2 |
| 3 | 2 | 0 | 103 | 0 | 0 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 93.6 |
| 4 | 4 | 4 | 0 | 35 | 10 | 3 | 7 | 13 | 5 | 1 | 0 | 1 | 5 | 39.8 |
| 5 | 0 | 1 | 1 | 0 | 64 | 0 | 0 | 24 | 8 | 0 | 0 | 2 | 4 | 61.5 |
| 6 | 7 | 1 | 19 | 1 | 0 | 77 | 17 | 4 | 0 | 0 | 0 | 0 | 0 | 61.1 |
| 7 | 1 | 10 | 42 | 3 | 1 | 17 | 57 | 5 | 1 | 0 | 0 | 0 | 1 | 41.3 |
| 8 | 1 | 0 | 1 | 0 | 10 | 5 | 3 | 64 | 11 | 0 | 0 | 0 | 3 | 65.3 |
| 9 | 10 | 0 | 5 | 0 | 4 | 8 | 1 | 14 | 63 | 4 | 0 | 1 | 10 | 52.5 |
| 10 | 2 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 4 | 36 | 0 | 1 | 25 | 49.3 |
| 11 | 3 | 0 | 0 | 0 | 5 | 2 | 2 | 7 | 4 | 7 | 7 | 4 | 31 | 9.7 |
| 12 | 1 | 0 | 0 | 0 | 7 | 0 | 0 | 5 | 3 | 4 | 3 | 12 | 35 | 17.1 |
| 13 | 7 | 0 | 0 | 0 | 6 | 3 | 0 | 3 | 10 | 5 | 0 | 1 | 62 | 63.9 |
| p(%) | 60.0 | 74.6 | 58.2 | 67.3 | 56.1 | 59.7 | 52.3 | 44.8 | 55.8 | 57.1 | 58.3 | 54.5 | 33.5 | |

TABLE 1 – La matrice de confusion avec la taille de vocabulaire 10000

À partir du résultat de cette expérimentation nous pouvons conclure que la taille du vocabulaire influence le taux de reconnaissance. En fait, dans notre cas, la taille de vocabulaire optimale est environ 2000. Le taux de reconnaissance plus grand obtenue est 56% avec la taille de vocabulaire de 2000 et les paramètres de SVM c=300 et gamma = 0.5. En voyant la figure 9, on constate que le temps d'apprentissage augmente linéairement quand la taille de vocabulaire augmente. Quand cette taille égale 10, le temps d'apprentissage est 154 seconds, quand la taille égale 10000, le temps est 6986 seconds. Cependant, le temps de reconnaissance augmente très lentement par rapport à l'augmentation de la taille de vocabulaire (figure 8). Par exemple, quand la taille de vocabulaire est 10, le temps de reconnaissance (une reconnaissance de 1290 images) est 79 seconds et celui quand la taille est 10000 est seulement 126 seconds. On peut conclure que quand la taille de vocabulaire augmente, le temps pour l'apprentissage augmente beaucoup mais le temps pour la reconnaissance augmente un peu.

## Analyses sur la matrice de confusion

Au vu la matrice de confusion (Table 1) on constate qu'il y a deux scènes que le programme ne reconnaît pas bien, ce sont la scène 11 et 12 (bedroom et kitchen), les taux de rappels sont 9.7

et 17.1%. Elles sont mal classifiées à la scène 13 (living room). On peut expliquer que dans ces images, il y a beaucoup d'objets ressembles tels que : table, chaise, lampe ...

Nous allons faire une autre expérimentation sans les scènes à intérieurs (sans scène 11, 12, 13). Le résultat de cette expérimentation est meilleur avec le taux de reconnaissance de 66%. Le tableau 2 est la matrice de confusion de cette expérimentation. Ce résultat montre que notre algorithme rend de meilleurs résultats sur les scènes à l'extérieurs, les résultats des scènes à l'intérieur sont moins bons.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Taux de rappel |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 76 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 1 | 95.0 |
| 2 | 7 | 65 | 2 | 18 | 1 | 9 | 12 | 2 | 0 | 4 | 54.2 |
| 3 | 1 | 0 | 95 | 0 | 0 | 4 | 9 | 0 | 0 | 0 | 87.2 |
| 4 | 6 | 13 | 1 | 39 | 10 | 1 | 5 | 6 | 3 | 2 | 45.3 |
| 5 | 1 | 0 | 1 | 0 | 71 | 0 | 1 | 15 | 9 | 4 | 69.6 |
| 6 | 5 | 4 | 8 | 1 | 0 | 81 | 22 | 2 | 0 | 1 | 65.3 |
| 7 | 7 | 9 | 32 | 2 | 1 | 19 | 63 | 2 | 1 | 0 | 46.3 |
| 8 | 1 | 1 | 1 | 0 | 13 | 3 | 8 | 60 | 10 | 0 | 61.9 |
| 9 | 11 | 0 | 2 | 0 | 9 | 2 | 4 | 16 | 70 | 4 | 59.3 |
| 10 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 1 | 66 | 93.0 |
| p | 65.5 | 70.7 | 66.9 | 65.0 | 66.4 | 68.1 | 50.0 | 57.7 | 73.7 | 80.5 | |

TABLE 2 – La matrice de confusion avec la taille de vocabulaire 1000, sans les scènes : living room, bedroom, kitchen

# 5 Conclusion

Dans ce domaine, il y a des méthodes qui permettent de classifier bien deux catégories mais pour plusieurs catégories, elles ne classifient pas bien, pour cela on perd de temps pour essayer. Durant ce TP, on a faire la recherche sur le domaine de classification des scènes naturelles, cherche à comprendre quelques méthodes tels que Pyramid Match Kernels, Bag of Visual Words, etc. Finalement, on a implémenté la méthode Bag of Visual Words. Bien qu'il y a des défis dans ce domaine, on ait réussi à implémenter une méthode qui sert à classifier les catégories d'images avec le résultat acceptable.
Nous avons expérimenté plusieurs scénarios. Le résultat de notre programme montre que la reconnaissance des scènes à l'intérieur est plus difficile que celle des scènes à l'extérieur.

# Bibliographie


[1] Jun Hartigan, John A., and Manchek A. Wong. *Algorithm AS 136 : A k-means clustering algorithm.*. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979) : 100-108.

[2] Burges, Christopher JC *A tutorial on support vector machines for pattern recognition.*. Data mining and knowledge discovery 2.2 (1998) : 121-167. APA

[3] Yang, Jun, et al. *Evaluating bag-of-visual-words representations in scene classification.*. Proceedings of the international workshop on Workshop on multimedia information retrieval. ACM, 2007.

[4] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, Chong-Wah Ngo *EvaluatingBag-of-Visual-Words RepresentationsClassification*.

[5] Svetlana Lazebnik, Cordelia Schmid, Jean Ponce. *Beyond Bags of Features : Spatial Pyramid Matching for Recognizing Natural Scene Categories*

[6] NGUYEN Thi Oanh 2013, *transparent sur Bag of Visual Words, IFI Hanoi*.