TP : Classification d'actions avec des descripteurs locaux

Univ. Lille – Master Informatique, parcours RVA, UE AIV2

1 Objectifs

On s'intéresse dans ce TP à la classification d'actions dans des vidéos. L'objectif est de mettre en place un pipeline de classification d'actions basé sur la détection de points d'intérêt spatio-temporels (STIP – *space-time interest points*), leur représentation à l'aide de descripteurs d'apparence (HOG) et de mouvement (HOF), la constitution de sacs de mots visuels spatio-temporels décrivant les vidéos complètes, et enfin leur classification à l'aide d'un classifieur supervisé.

2 Données

Les jeux de données suivants seront utilisés :

— `ucf-sports` : un ensemble de 150 vidéos de sport représentant dix classes d'actions distinctes (*Diving, Golf, Kicking, Lifting, Riding, Run, SkateBoarding, SwingBench, SwingSide, Walk*) ;

— `ucf-sports_augmented` : le même jeu de données dans lequel les vidéos ont été inversées horizontalement. Ce jeu de données servira à augmenter le jeu de données d'entraînement pour accroître la robustesse du classifieur.

Les jeux de données sont disponibles dans l'archive `ucf-sports_dataset.tar.gz` accessible depuis la page Moodle du cours. L'archive contient :

— un répertoire `videos` qui contient les vidéos des deux jeux de données au format `avi` ;

— un répertoire `keypoints` qui contient pour chaque vidéo les points d'intérêt extraits de cette vidéo et les descripteurs locaux HOG et HOF extraits au voisinage de ce point. Les fichiers `.key` sont des fichiers texte dont chaque ligne contient les données d'un point d'intérêt et les descripteurs associés ; la fonction de lecture de ces fichiers et la signification des données sont décrites en Section 3 ;

— un fichier texte `ucf-sports.files` qui contient la liste des fichiers du jeu de données UCF-Sports et les labels associés (chaque ligne contient un nom de fichier et le label associé) ;

— un fichier `ucf-sports_augmented.files` qui contient la liste des fichiers du jeu de données augmenté, au même format que le fichier `ucf-sports.files`.

Enfin, des vocabulaires visuels sont fournis dans l'archive `visual_vocabularies.tar.gz`. Ces vocabulaires seront utilisés pour le calcul des sacs de mots visuels des vidéos. Chaque vocabulaire est une matrice de taille `(taille_du_vocabulaire, dimension_des_descripteurs)` stockée dans un fichier au format standard Numpy. Le nommage des fichiers suit la convention suivante : `voc_d_n.npy`, où `d` désigne le type de descripteurs (`hog`, `hof` ou `hoghof`) et `n` désigne la taille du vocabulaire visuel.

Note : Les fichiers `.files` contiennent uniquement les noms de fichiers, sans chemin ni extension. Il vous appartient de les compléter pour accéder aux fichiers `.avi` ou `.key` correspondants.

3 Code

Un module `stip` (fichier `stip.py`) est fourni sur la page Moodle du cours. Il contient une fonction permettant de lire les fichiers `.key` contenant les points d'intérêt spatio-temporels

(STIP) et les descripteur (HOG et HOF) associés. La fonction retourne un tuple `(keypoints, descriptors)`, tel que :

— `keypoints` est une matrice contenant les points d'intérêt détectés (un point par ligne). Chaque point suit le format suivant :
`y x t scale_xy scale_t`
avec `(x,y)` la position spatiale du point dans sa trame, `t` l'indice de la trame dans laquelle a été détecté le point d'intérêt, `scale_xy` l'échelle spatiale du point et `scale_t` son échelle temporelle ;

— `descriptors` est une matrice contenant les descripteurs associés aux points d'intérêt. Chaque ligne contient un descripteur de dimension 162 : les 72 premières composantes correspondent à l'histogramme de gradient (HOG) dans le voisinage du point d'intérêt ; les 90 composantes suivantes correspondent à l'histogramme de flux optique (HOF) calculé dans le voisinage du point d'intérêt.

4 Visualisation des points d'intérêt

Dans un premier temps, nous allons uniquement visualiser les points d'intérêt STIP sur un échantillon de vidéos.

Questions

1. Charger les points d'intérêt et descripteurs des vidéos du jeu de données `ucf-sports`.
2. Écrire une fonction pour visualiser les points d'intérêt détectés dans une vidéo. Observez sur quelques vidéos la nature des points détectés.

5 Classification par sacs de mots visuels spatio-temporels

Nous allons maintenant mettre en place un système de classification basé sur les descripteurs locaux calculés au voisinage des points d'intérêt. Les descripteurs locaux seront agrégés via la méthode des sacs de mots visuels. Trois descripteurs locaux seront comparés : HOG, HOF, et la concaténation des deux (HOG+HOF).

Questions

1. Écrire une fonction pour calculer le vecteur de fréquences du sac de mots visuels correspondant à une vidéo à partir de ses descripteurs locaux HOG+HOF et du vocabulaire visuel correspondant.
2. Écrire une fonction pour calculer ces vecteurs pour l'ensemble des vidéos du jeu de données.
3. Réaliser le travail équivalent pour établir des vecteurs de fréquences de mots visuels à partir :
   (a) des descripteurs HOG (72 premières composantes des descripteurs) uniquement ;
   (b) des descripteurs HOF (90 dernières composantes des descripteurs) uniquement.
4. Mettre en place le processus de classification supervisée, selon le protocole suivant :
   — utiliser un classifieur SVM multiclasse [1].
   — utiliser un protocole *leave-one-out* [2], car les données sont disponibles en faible quantité, pour entraîner et tester votre classifieur.
   — calculer le taux de classification (*accuracy*) moyen de votre classifieur sur l'ensemble des apprentissages réalisés en leave-one-out.
5. Comparez les performances en classification obtenues avec les différents descripteurs (HOG+HOF, HOG, HOF). Comment interpréter ces résultats du point de vue de l'information visuelle portée par les descripteurs ?

1. Dans `scikit-learn` : `classe sklearn.svm.SVC`.
2. Dans `scikit-learn` : `classe sklearn.model_selection.LeaveOneOut`.

6 Augmentation de données

Pour améliorer la qualité du classifieur, il est possible d'utiliser les données augmentées du jeu de données `ucf-sports_augmented` pour avoir un jeu d'entraînement de plus grande taille. On utilisera donc un ensemble d'apprentissage contenant à la fois les vidéos d'apprentissage du jeu de données d'origine et les vidéos équivalentes (versions inversées horizontalement) du jeu d'augmentation. Seul l'ensemble d'apprentissage sera augmenté ; l'ensemble de test restera réduit à une vidéo (*leave-one-out*) du jeu de données initial. Attention : il ne faut pas ajouter au jeu d'apprentissage la version augmentée de la donnée de test, car cela induirait un biais dans les résultats.

Questions

1. Modifiez votre protocole d'apprentissage pour qu'à chaque entraînement du classifieur, le jeu d'entraînement comprenne les vidéos d'entraînement et leur équivalent dans le jeu de données augmentées.
2. Reproduisez les expérimentations de la section précédente en utilisant ce nouveau protocole.
3. Interprétez les résultats obtenus.