Un environnement efficace pour la classification d'images à grande échelle

Thanh-Nghi Doan*, François Poulet*,**

**Université de Rennes I, *IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France,
{thanh-nghi.doan | francois.poulet}@irisa.fr

Résumé. La plupart des processus de classification d'images comportent trois principales étapes : l'extraction de descripteurs de bas niveaux, la création d'un vocabulaire visuel par quantification et l'apprentissage à l'aide d'un algorithme de classification (eg.SVM). De nombreux problèmes se posent pour le passage à l'échelle comme avec l'ensemble de données ImageNet contenant 14 millions d'images et 21,841 classes. La complexité concerne le temps d'exécution de chaque tâche et les besoins en mémoire et disque (eg. le stockage des SIFTs nécessite 11To). Nous présentons une version parallèle de LibSVM pour traiter de grands ensembles de données dans un temps raisonnable. De plus, il y a beaucoup de perte d'information lors de la phase de quantification et les mots visuels obtenus ne sont pas assez discriminants pour de grands ensembles d'images. Nous proposons d'utiliser plusieurs descripteurs simultanément pour améliorer la précision de la classification sur de grands ensembles d'images. Nous présentons nos premiers résultats sur les 10 plus grandes classes (24,817 images) d'ImageNet.

1 Introduction

La classification d'images est une tâche importante dans le domaine de la vision par ordinateur, la reconnaissance d'objets et l'apprentissage automatique. L'utilisation de descripteurs de bas niveau de l'image et le modèle de sac de mots sont au coeur des systèmes de classification d'images actuels. La plupart des environnements de classification d'images comportent trois étapes : 1) l'extraction de descripteurs de bas niveau dans les images, 2) la création d'un vocabulaire de mots-visuels et 3) l'apprentissage du modèle des classes d'images. Dans la première étape d'extraction des descripteurs de bas niveau, les choix les plus courants dans les méthodes récentes sont les SIFTs (Lowe (2004)), les SURFs (Bay et al. (2006)) ou les DSIFTs (Bosch et al. (2007)). L'étape 2 est la création du vocabulaire visuel, le choix habituel pour cette étape est l'utilisation d'un algorithme de k-means et la création d'un sac de mots. La troisième étape est l'apprentissage du classifieur, beaucoup de systèmes choisissent souvent des Support Vector Machines avec des noyaux linéaires ou non-linéaires. La plupart de ces systèmes sont ensuite évalués sur de petits ensembles de données qui tiennent sans problème en mémoire centrale, comme Caltech-101 (Fei-Fei et al. (2004)), Caltech-256 (Griffin et al. (2007)) ou PASCAL VOC (Everingham et al. (2010)). Cependant l'apparition notamment de l'ensemble