

Création de Vocabulaires Visuels Efficaces pour la Catégorisation d'Images.*

Creating Efficient Visual Codebooks for Image Categorization.

Diane LARLUS

Gyuri DORKÓ

Frédéric JURIE

GRAVIR - CNRS - INRIA Rhône-Alpes - INPG - UJF - France
 {diane.larlus,gyuri.dorko,frederic.jurie}@inrialpes.fr

Résumé

Nous proposons dans cet article une méthode de construction automatique de vocabulaires visuels. Le vocabulaire visuel est obtenu par quantification de descripteurs locaux des images.

Les vocabulaires visuels produits sont utilisés pour construire automatiquement des représentations discriminantes des objets présents dans les images.

Nous décrivons une application de ces techniques à la catégorisation d'images par sacs de primitives (bags of features) et montrons que les résultats obtenus sont très supérieurs à ceux obtenus par les méthodes concurrentes.

Mots Clef

représentation des images, catégorisation d'image, apprentissage automatique.

Abstract

We propose in this article an automatic method for building visual codebooks. Codebooks are obtained by quantizing local image descriptors and are used to automatically build discriminative representations of objects occurring in images.

We describe an image categorization application based on the proposed approaches, providing results far above related state of the art existing methods.

Keywords

image description, image categorization, machine learning

1 Introduction

Appareils photo numériques, photophones, caméscopes, caméras de surveillance, etc., l'image numérique est désormais omniprésente dans notre société. Si le stockage et la transmission des images numériques semblent être un problème en voie d'être résolu, celui de leur interprétation automatique demeure un problème difficile. Or, l'interprétation automatique des images ouvrirait la porte à de nombreuses applications pour lesquelles la machine



Figure 1: Y a-t-il une bicyclette dans l'image ? Peut-on construire des algorithmes capables de répondre à cette question ?

viendrait au secours de l'homme, incapable de traiter de telles masses d'images. Nous pouvons par exemple imaginer des systèmes automatiques capables d'analyser simultanément plusieurs milliers de caméras de surveillance, ou encore un assistant qui viendrait aider l'utilisateur à rechercher des images dans une collection de photos trop importante pour être parcourue entièrement.

Parmi les différents outils d'analyse automatique des images, nous nous intéressons en particulier à ceux permettant de catégoriser (*ie* classer) des images en fonction de la présence/absence d'objets particuliers. Cet outil doit être capable de gérer les variations de point de vue, d'illumination, les occultations, et les variations de forme à l'intérieur même d'une classe d'objet.

Les méthodes développées dans notre travail ont recours à des techniques d'apprentissage : l'utilisateur fournit au système des images contenant les objets d'intérêt, et ces images sont utilisées pour construire automatiquement des modèles des catégories d'objets ainsi que des règles de classification (illustration figure 1).

Bien que seul le contexte de la classification d'image soit abordé ici, les techniques développées peuvent également s'appliquer à la détection d'objets. Dans ce cas, l'algorithme est appliqué à des fenêtres d'analyse et une décision de présence/absence d'objets est prise pour chacune de ces fenêtres. Ces fenêtres peuvent provenir d'une exploration systématique de l'image (à toutes les échelles et à toutes les positions) ou de prétraitements tels que l'extraction de zones en mouvement.

*Travail réalisé avec le support de la société THALES Optronique et du réseau d'excellence PASCAL

1.1 Représentation locale des images

Contrairement à d'autres types d'informations (tels que les textes écrits par exemple), les informations portées par les structures de base de l'image ne sont pas assez riches pour pouvoir être utilisées individuellement. Un pixel ne contient qu'un niveau de gris (ou un triplet de valeurs RGB dans le cas d'images couleurs) ce qui ne permet pas, en général, d'inférer directement des informations sur le contenu de l'image.

C'est pourquoi les pixels sont généralement utilisés non pas seuls, mais par l'intermédiaire d'*indices visuels*, qui sont des groupes de pixels combinés au sein d'une même structure. Depuis l'origine de la vision par ordinateur de très nombreux types d'indices visuels ont été proposés et utilisés pour interpréter le contenu des images. Nous pouvons citer, par exemple, les contours, les régions homogènes, les descripteurs locaux de textures, les jets locaux, les textons, etc. [8].

Dans notre cas, nous utilisons une représentation locale des images au moyen du descripteur SIFT (Scale Invariant Feature Transform [19]). Ce descripteur a été utilisé avec succès dans de nombreuses applications de reconnaissance d'objets et permet d'obtenir les meilleures performances dans un contexte de mise en correspondance de points d'intérêt [22]. En un point particulier de l'image, le descripteur est constitué d'un histogramme grossier des orientations des gradients contenus dans un voisinage de ce point. De manière plus précise, les gradients sont calculés selon 8 orientations différentes, dans chaque case d'une grille 4x4 centrée sur ce point particulier, donnant ainsi des vecteurs de dimension 128¹. La taille de la fenêtre utilisée pour définir le voisinage est directement liée à l'échelle à laquelle l'image est analysée.

Ainsi, dans tous les travaux décrits dans cet article, les images ne sont jamais utilisées au moyen de leurs niveaux de gris. Toutes les images manipulées sont d'abord transformées en représentation SIFT multi-échelle.

1.2 Représentation des images et stratégie de catégorisation

Pour représenter globalement une image ou un objet, nous avons adopté une approche de type *bag of features*. Il s'agit d'une approche initialement développée pour la catégorisation de documents écrits, domaine dans lequel elle s'est avérée très performante [14]. Un vocabulaire de mots (sélectionné de manière pertinente) est dans un premier temps constitué. Dans le cas du texte il s'agit de repérer les mots utiles pour discriminer les catégories de documents. Chaque document est alors représenté par un histogramme représentant la fréquence de chaque mot du vocabulaire. Les histogrammes subissent diverses normalisations visant

¹Ces valeurs numériques sont celles utilisées dans cet article. Elles peuvent bien entendu être modifiées pour s'adapter à différents contextes. Nous avons repris les valeurs préconisées par D. Lowe, valeurs déterminées de manière à donner des performances optimales de mise en correspondance de points d'intérêt [19].

à les affranchir de la taille du document et à amplifier les mots discriminants.

Différentes stratégies peuvent être ensuite utilisées pour la classification des documents. Nous notons en particulier les deux approches les plus typiques que sont la classification par MAP ou ML [12] impliquant des modèles génératifs des documents (construits à partir d'exemples) ou la classification par recherche de fonctions discriminantes (construites à partir d'exemples) par exemple au moyen de classifieurs SVM ou K-PPV [12].

Ce type d'approche par *bag of features* peut être transposé au cas de la catégorisation d'images [4]. Dans ce cas, les images sont caractérisées par un histogramme qui compte le nombre d'occurrences de chaque type de représentations locales. Par analogie, ces représentations locales particulières seront dénommées par la suite *vocabulaire* ou *vocabulaire visuel*.

Bien entendu, contrairement au cas de l'analyse de documents, le vocabulaire visuel n'est pas une donnée intrinsèque aux images. Il n'existe pas de vocabulaire unique pour décrire les images. Ce vocabulaire doit donc être construit pour répondre à des propriétés particulières.

C'est dans ce point que réside le coeur de nos travaux : comment produire le *meilleur* vocabulaire, c'est-à-dire celui qui permettra d'obtenir les meilleures performances de catégorisation d'images.

Une fois le vocabulaire construit, nous représentons l'image en prenant tous les descripteurs se trouvant sur les noeuds d'une grille régulière (à la fois en position et échelle) et en remplaçant le descripteur par le mot du vocabulaire qu'il représente (s'il représente un mot du vocabulaire ; nous verrons que certaines structures peuvent ne pas être affectées à des mots). Les techniques utilisées pour le texte deviennent alors applicables.

Nous utiliserons en particulier ici une classification au moyen de classifieurs binaires SVM linéaires [23], choisis en raison de leurs bonnes performances même dans le cas de problèmes de haute dimensionnalité².

Nous entraînons un classifieur par catégorie d'objets, pour chaque catégorie dont la présence doit être détectée dans l'image.

1.3 Construction de vocabulaires visuels

Comme nous venons de le dire, la construction du vocabulaire est au coeur de notre étude. Aucun vocabulaire visuel n'existant de manière implicite, il s'agit de s'interroger sur les propriétés que doivent posséder les vocabulaires visuels et sur les méthodes à utiliser pour les construire.

La construction du vocabulaire suppose une quantification de l'espace de représentation des descriptions locales des images. Il s'agit en effet de construire une fonction de l'espace de représentation (SIFT : R^{128} dans notre cas) vers un espace discret de labels.

²D'autres classifieurs ont été testés mais ces expériences ne sont pas reportées dans cet article

L'espace de représentation locale des images n'est pas peuplé de manière dense et uniforme. Certains motifs visuels (théoriquement possibles) peuvent ne jamais apparaître dans les images tandis que d'autres peuvent être très fréquents. La première conséquence de cette remarque est que le vocabulaire doit être adapté aux images rencontrées, c'est-à-dire il doit être le reflet des descriptions locales présentes dans les images.

La méthode la plus utilisée pour construire un vocabulaire visuel consiste à partir des descripteurs rencontrés dans les images (statistiquement représentatives du problème) et de les regrouper en un nombre fini de *clusters* au moyen d'un algorithme de *clustering*. Ce nombre de clusters représente la taille du vocabulaire.

2 Création de vocabulaires visuels par clustering de descripteurs locaux

La création de *vocabulaires visuels* basée sur la quantification de descripteurs d'apparences visuels constituent un moyen efficace de capturer les statistiques de l'image, que cela soit pour faire de la reconnaissance d'objets ou de l'analyse de textures.

La création du vocabulaire est liée à trois types de composants : la description locale des images, le choix des parties de l'image utilisées (par exemple certaines méthodes ne traitent l'image qu'en un nombre limité de points d'intérêts alors que d'autres traitent tous les points de l'image), l'algorithme utilisé pour la quantification. Nous nous concentrons dans cet article sur la méthode utilisée pour la quantification.

Dans la majorité des cas décrits dans la littérature, le vocabulaire est créé à partir d'un ensemble d'images d'apprentissage au moyen d'un algorithme de clustering.

Le terme de *texton*, proposé initialement par Julesz il y a 20 ans de cela, représente un ensemble de *patches*³ représentatifs des apparences locales des images. D'une manière plus générale, les textons représentent des micro-structures élémentaires de l'image [9, 18, 27].

Dans [18], Leung et Malik proposent une méthode de construction de texton. L'idée consiste à représenter localement l'image au moyen de convolutions avec des banques de filtres de Gaussiennes orientés, puis à quantifier ces représentations locales avec un algorithme *k-means*.

Ce type de méthode a été étendu par Hall et Crowley [11]. Plutôt que de ne considérer les textons qu'à une seule échelle, ces auteurs proposent de déterminer l'échelle intrinsèque en chaque point de l'image, et d'utiliser cette échelle dans la construction du vocabulaire. La quantification est réalisée par *k-means*. Ici encore, l'ensemble de l'image est décrite par des mots du vocabulaire, et des histogrammes permettent la reconnaissance des visages.

³Un patche désigne une partie de l'image dont le support spatial est très limité

Lorsque les structures de l'image représentée par le vocabulaire ont une taille importante dans l'image, les méthodes éparées, où toute l'image n'est pas utilisée, sont généralement préférées.

Une des premières approches utilisant un vocabulaire éparse est celle proposée par Weber *et al.* [25]. Le vocabulaire est appris à partir d'un ensemble d'images dont des points d'intérêts sont extraits avec un détecteur de Förstner, à une seule échelle. La quantification de ce petit nombre de patches est réalisée par *k-means*.

Plus récemment, Leibe *et al* [17] ont proposé une méthode de détection basée sur une stratégie de vote dans un espace de transformation. Ils utilisent un vocabulaire appris en détectant des points de Harris dans des images d'apprentissage, puis en regroupant ces patches en groupes au moyen d'un algorithme de clustering hiérarchique. Les descripteurs locaux sont des vecteurs de niveaux de gris.

Le détecteur proposé par Agarwal *et al* [1] repose sur une approche similaire, mais en incorporant des relations géométriques entre les mots du vocabulaire. Ce vocabulaire est créé de la même manière que dans [17].

Csurka *et al* [4] proposent une approche de type *bag of features*, pour laquelle les descripteurs SIFT [19] sont utilisés pour représenter localement l'image. L'image n'est utilisée qu'en un nombre réduit de points, sélectionnés par l'algorithme Harris-affine [21]. La quantification conduisant au vocabulaire est obtenue par un clustering des descripteurs locaux par *k-means*.

3 Vers un nouvel algorithme de création de vocabulaires

Nous avons vu dans la section précédente que de nombreuses méthodes ont été proposées pour créer des vocabulaires visuels à partir d'ensembles de représentations locales d'images types.

Nous allons désormais présenter les raisons qui nous ont poussés à proposer une nouvelle méthode. Les méthodes rencontrées dans la littérature peuvent se regrouper principalement en deux catégories : celles basées sur un clustering de type *k-means*, celles basées sur un clustering de type agglomératif.

3.1 Spécificités et limitations

Méthodes de type k-means. Ces méthodes présentent une faiblesse majeure, comme souligné dans [10, 3, 16], qui est de ne pas être adaptées au cas de données mal équilibrées. Si certains clusters sont très denses par rapport à d'autres, *k-means* leur donnera une importance supérieure. Au bout du compte, certains clusters de faible densité risquent de ne pas être découverts et agglomérés à de gros clusters, même si ces derniers sont très distants. Or il a été montré que la densité des descripteurs locaux d'images n'est pas du tout uniforme [3, 16] et qu'ainsi les données à regrouper vont présenter ce caractère "non-équilibré".

De même, il est montré que dans le cas des images, ce ne sont justement pas les descripteurs les plus nombreux

qui sont les plus informatifs [24, 16]. Il faut donc que la méthode de production de vocabulaire prenne en compte cette particularité des données.

Ce type de problème est également connu en analyse de données sous le nom d'*analyse de cas rares* [26].

Un autre problème, certes moins bloquant car des solutions existent, est celui du nombre de clusters, ce dernier devant être fixé à l'avance.

Méthodes de type agglomératif. Nous ne représentons pas les images au moyen d'un petit nombre de points d'intérêt [1, 5, 7, 13, 17, 25] mais au moyen d'un échantillonnage dense et régulier de l'image. Des travaux [15] ont en effet récemment montré que si les points d'intérêt étaient très performants pour la mise en correspondance à partir d'images du même objet vu de différents points de vues, ils l'étaient moins pour l'appariement entre objets différents d'une même catégorie.

La conséquence de cet échantillonnage dense est l'accroissement du nombre de descripteurs obtenus par image. Là où une méthode par point d'intérêt retenait quelques centaines de points par image, nous en produisons plusieurs dizaines de milliers. Pour une base d'images typique, la production du vocabulaire repose alors sur le clustering de dizaines de millions de descripteurs. Même en adoptant des structures de données adaptées à la recherche rapide de plus proches voisins (structures arborescentes), le clustering agglomératif n'est pas applicable compte tenu du nombre de vecteurs.

Utiliser un échantillonnage uniforme des vecteurs pour en réduire le nombre n'est pas satisfaisant, pour les raisons évoquées ci-dessus : compte tenu du déséquilibre important des clusters, seuls les plus denses seraient échantillonnés correctement.

Notons que les méthodes agglomératives sont plus satisfaisantes que k-means quant à la définition du nombre de clusters. Avec les méthodes agglomératives, le nombre de cluster peut être adapté "en ligne" en fonction de données présentes, en imposant un seuil sur la distance maximale des vecteurs d'un même cluster, ce qui est plus intuitif que de fixer a priori le nombre de catégories.

3.2 Algorithme proposé

La méthode de clustering utilisée doit permettre : (a) de découvrir des clusters peu denses, (b) d'être adaptée au traitement de très grosses quantités de vecteurs, (c) de permettre une plus grande souplesse dans le choix du nombre de clusters

Nous avons retenu, pour faire face à ces exigences, les principes suivants :

- utiliser un algorithme de clustering travaillant sur des données échantillonnées (ne pas utiliser toutes les données en même temps)
- utiliser un échantillonnage biaisé : l'échantillonnage, s'il est uniforme va donner plus de poids aux clus-

ters les plus denses. Nous proposons d'introduire un échantillonnage biaisé qui compense cet effet.

- un clustering online : les clusters sont ajoutés un par un. A chaque nouvelle itération les centres de clusters déjà produits ne sont pas remis en cause.

En pratique nous proposons de combiner l'algorithme online median [20] avec un échantillonnage biaisé des données. Cet échantillonnage n'utilise pas de calcul d'estimation des densités. En revanche, les régions contenant les centres des clusters sont considérées comme denses et surreprésentées. L'échantillonnage est fait de façon à éviter ces zones.

Nous tirons parti du fait que les centres soient placés les uns après les autres pour alterner phases d'échantillonnage biaisé, et placement de nouveaux centres.

Online Median. L'algorithme *online median* [20], proposé par R. Mettu et C. Plaxton est une solution *online*⁴ du problème *k-median*. Au lieu d'optimiser globalement le placement de k centres (k fixé), ceux-ci sont placés un par un, selon le principe des *Facility Location*.

L'algorithme *online median* place successivement, et de façon définitive, les centres du clustering. Le processus s'arrête lorsqu'un critère d'arrêt est vérifié, ou lorsque toutes les données ont été choisies comme centre.

L'algorithme dispose de paramètres fixés $\alpha, \beta, \gamma, \delta$ propres à l'algorithme. Définissons tout d'abord la valeur d'une boule A de centre x et de rayon r .

$$val(A) = \sum_{y \in A} (r - d(x, y))$$

Le fils d'une boule A de centre x et de rayon r est un point y qui vérifie $d(x, y) < r\beta$. Rappelons que $d(y, X)$ représente $\min_{x \in X} d(y, x)$

Chaque étape d'ajout d'un centre se fait ainsi :

- (1) Calcul de la valeur de toutes les boules centrées en un point x des données D et de rayon $\frac{d(x, Z)}{\gamma}$ où Z est l'ensemble des centres déjà placés. Si $Z = \emptyset$ (cas du premier centre), le rayon choisi est $\max_{x, y \in D} d(x, y)$.
- (2) Sélection de la boule A_0 de valeur maximale.
- (3) Tant que A_i contient plusieurs fils
 - Considérons les boules centrées en y vérifiant $d(x_i, y) \leq \beta r_i$, de rayon $r_{i+1} = \frac{r_i}{\alpha}$. Soit A_{i+1} la boule de valeur maximale, et x_{i+1} le centre correspondant. Son rayon est le r_{i+1} précédent.
- Lorsque la boule A_i n'a qu'un seul fils, on le choisit comme nouveau centre du cluster.

⁴Un algorithme de clustering *online* au sens où nous l'entendons ici, place les centres un par un mais peut revenir autant de fois que nécessaire sur les données. Nous distinguons ce terme de *streaming* qui signifie que les données ne peuvent être vues qu'une fois, dans un ordre déterminé

Cela revient à mettre à jour une boule dont le rayon diminue à chaque itération (par $r_{i+1} = \frac{r_i}{\alpha}$) et qui se déplace à chaque étape vers la région de “plus forte densité”, estimée à travers la valeur.

Un clustering à deux phases. Comme nous venons de le signaler, nous utilisons le fait que les centres soient placés itérativement pour intercaler phases de rééchantillonnage (renouvellement des vecteurs traités) et phases de placement des centres, en utilisant les centres déjà placés pour guider l'échantillonnage.

Échantillonnage biaisé. L'idée retenue est de favoriser la découverte de nouvelles régions, au détriment des clusters très denses. Il faut donc échantillonner loin des centres déjà placés. Pour cela un rayon d'influence est défini. Tous les vecteurs contenus dans une boule centrée sur un centre déjà placé et de rayon ce rayon d'influence sont considérés comme affectés au cluster correspondant et ne seront pas échantillonnés.

Lors d'une étape d'échantillonnage, on choisit uniquement des points hors des boules d'influence de tous les centres déjà placés. La sphère d'influence est un paramètre de l'algorithme dont nous étudierons l'influence dans la section 4. Plutôt que de considérer que les régions d'influence sont des boules, un modèle probabiliste pourrait être utilisé (modèle Gaussien par exemple).

Le rôle de cette boule d'influence est de limiter l'effet des zones de très fortes densités afin qu'un centre ne soit pas placé au même endroit qu'un centre précédent. En effet, ceci permet d'éviter les redondances dans le vocabulaire, et de trouver des classes moins peuplées qui peuvent s'avérer discriminative.

Utilisation du vocabulaire, construction des histogrammes. Le vocabulaire est constitué de l'ensemble des centres produits par l'algorithme que nous venons de décrire, à partir d'un ensemble d'apprentissage.

Comme expliqué section 1.2, chaque descripteur local d'une image doit être traduit en mot de vocabulaire.

Nous avons imaginé deux règles de traduction correspondant à deux visions différentes du problème : (a) affectation du descripteur au cluster dont le centre est le plus proche du descripteur, quelque soit la distance qui les sépare, (b) affectation du descripteur à l'ensemble des clusters pour lesquels le point tombe dans la zone d'influence.

Dans le premier cas, nous cherchons à quantifier tout l'espace des descripteurs à l'aide des éléments du vocabulaire. L'espace est ainsi entièrement partitionné.

Dans le deuxième cas, de même que pour la phase de clustering. Les points sont considérés comme appartenant à un cluster s'ils sont dans la boule d'influence centrée en son représentant. Le reste est considéré comme du bruit.

4 Expérimentations

Nous décrivons dans cette section un ensemble d'expérimentations dont l'objet est de valider nos



Figure 2: Exemples d'images d'apprentissage (1ère ligne) et de test (2ème ligne) pour les quatre classes d'images.

idées, sur un problème de catégorisation d'images difficile et de taille réelle.

La majorité des expériences présentées ici sont liées aux données proposées dans le cadre de la compétition *The PASCAL Visual Object Classes Challenge*⁵ [6]. Nous présentons dans un premier temps ces données ainsi que la méthodologie d'évaluation, suivi des résultats obtenus par notre méthode montrant l'influence d'un certain nombre de facteurs. Cette partie est complétée par des expériences menées sur la base TUGraz.

4.1 The PASCAL Visual Object Classes Challenge

Le but de cette compétition est la reconnaissance d'objets appartenant à un certain nombre de classes, dans des scènes réalistes (*i.e.* non pré-segmentées). Un ensemble d'apprentissage et un ensemble de test, avec des images étiquetées sont fournis. Ils contiennent chacun 4 classes : (1) les vélos (provenant de la base TUGraz), (2) les voitures (provenant des bases CalTech, Leibe, Graz, Agarwal&Roth), (3) les motos (provenant des bases CalTech et ETHZ) et (4) les personnes (provenant de la base TUGraz). La figure 2 montre quelques images de chaque catégorie. Le nombre d'images d'apprentissage et de test pour les différentes catégories sont respectivement : Vélo 114/114, Voiture 272/275; Moto 214/216, Personne 84/84. L'ensemble d'apprentissage contient un ensemble d'images pour lesquelles le nombre, la position, et la classe des objets présents sont connus. Nous nous intéresserons ici à la classification : pour chacune des 4 classes, prédire la présence ou l'absence d'une instance de l'objet dans l'image de test.

L'évaluation des résultats se fait à l'aide de courbes ROC (une par catégorie), ainsi que par comparaison d'un taux de vrais positifs associé pour l'*Equal Error Rate*.

La courbe ROC représente la valeur du taux de vrais positifs (VP) en fonction du taux de faux positifs (FP), c'est à dire du taux de détection correcte d'éléments appartenant à la classe en fonction du taux d'éléments détectés par erreur comme appartenant à la classe.

Le taux de vrais positifs pour l'*Equal Error Rate* correspond au taux de vrais positifs obtenu pour le point de fonctionnement pour lequel $VP = 1 - FP$.

⁵<http://www.pascal-network.org/challenges/VOC/voc/index.html>

4.2 Catégorisation d'images à partir de vocabulaires visuels

Nous avons combiné la méthode de génération de vocabulaire décrite dans la section 3 avec la méthode de classification par *bag of features* présentée dans l'introduction (section 1.2).

Rappelons que cette méthode consiste à échantillonner régulièrement, de manière multi-échelle les images et à transformer chaque descripteur local en mot du vocabulaire. A chaque image correspond donc un histogramme normalisé représentant la fréquence de chaque mot du vocabulaire dans l'image.

Un classifieur linéaire SVM est entraîné sur les images d'apprentissage, dont les classes sont connues (la présence des objets est connue).

Nous utilisons des paramètres pour la grille d'échantillonnage des images qui donne environ 11000 descripteurs par image, ce qui donne au total plus de 8 millions de vecteurs. Ce choix, arbitraire, permet d'obtenir un échantillonnage relativement dense des images. Le pas d'échantillonnage est adapté aux échelles.

Les résultats obtenus (quantifiés par des courbes ROC) dépendent des facteurs suivants :

1. Valeur du rayon d'influence dans le clustering : ce rayon contrôle le nombre d'éléments affectés à chaque centre.
2. Utilisation ou non d'un rayon d'influence pour la construction des histogrammes : l'affectation à un élément du vocabulaire ne se fait qu'à l'intérieur de ce rayon.
3. Nombre de mots du vocabulaire : le vocabulaire utilisé peut être plus ou moins complet.
4. Nombre de centres ajoutés à chaque itération de l'algorithme : sur un même échantillonnage des données, un certain nombre d'éléments du vocabulaire sont produits.
5. Nombre d'échantillons sélectionnés à chaque itération : l'algorithme utilise un échantillonnage plus ou moins important des données.
6. Normalisation de l'histogramme : le classifieur utilise une version normée de la représentation des images, différentes normalisations sont proposées.
7. Données utilisées pour la création du vocabulaire : utilisation de toute l'image ou uniquement de la partie des images contenant l'objet.

L'étude jointe des influences des paramètres est difficilement réalisable, compte tenu du nombre de combinaisons possibles ; nous avons donc cherché à mettre en avant les paramètres les plus influents, que nous supposons fixés pour les expériences suivantes.

Nous avons également comparé la méthode proposée avec : une méthode n'utilisant pas d'échantillonnage biaisé, une méthode consistant à sélectionner aléatoirement les mots du vocabulaire parmi les descripteurs, et enfin une méthode utilisant un vocabulaire obtenu par l'algorithme *k-means*.

Normalisation. Les histogrammes obtenus doivent être normalisés de manière à les rendre invariants à différents

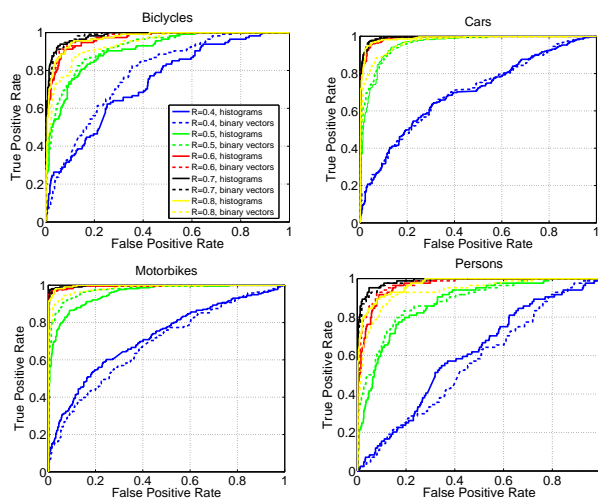


Figure 3: Effet du rayon d'influence sur les performances de l'algorithme.

paramètres, comme en particulier la taille des images.

Différents types de normalisation sont possibles. Nous avons comparé trois types de normalisation. La première consiste à centrer et normer le nombre de chacun des mots pour l'ensemble des images (normalisation par mot). La deuxième centre et norme les vecteurs (normalisation par image). La troisième binarise les vecteurs en seuillant à 1 toutes les valeurs différentes de 0.

Ces trois types de normalisation donnent, sur cette base de test, approximativement les mêmes résultats. Les expériences présentées ci-dessous sont obtenues avec une normalisation par mot.

Influence du rayon. Nous avons mesuré l'évolution des performances en fonction du rayon d'influence (définition section 3.2), à la fois dans le clustering, et dans la création des histogrammes.

Avec un rayon trop faible, même 4000 centres ne suffisent pas à avoir une bonne représentation des images. Il faut donc un rayon d'au moins 0.6, valeur utilisée par la suite, pour tester les autres paramètres. Avec un rayon trop élevé, on impose une trop grande distance entre les centres, et les points sont très vite tous affectés. Les derniers centres trouvés ne correspondent qu'à du bruit.

Il serait intéressant de mesurer également l'influence de ce rayon quand il n'est utilisé que pour le clustering, et que les histogrammes sont toujours calculés par affectation au centre le plus proche.

En conclusion, comme le montrent les graphiques de la figure 3, le rayon d'influence a un effet important sur les performances. Nous l'avons par la suite fixé à une valeur constante de 0.6, valeur numérique typiquement utilisée pour comparer des descripteurs SIFT dans un contexte de mise en correspondance. Il serait préférable de fixer cette valeur lors d'une phase de validation.

Utilisation du rayon d'influence pour le calcul des histogrammes. Pour réaliser l'histogramme, il est bien

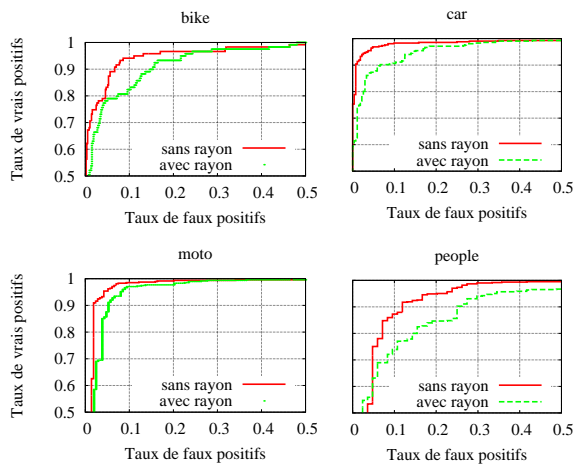


Figure 4: Comparaison pour le même vocabulaire visuel, des résultats obtenus avec une affectation au plus proche et l'utilisation du rayon d'influence utilisé dans le clustering, lors de la construction des histogrammes.

meilleur d'utiliser une affectation totale, c'est-à-dire de créer une partition des descripteurs, par affectation de tous les points à l'élément du vocabulaire le plus proche, que de garder le rayon d'influence du clustering (classification avec rejet), comme le montre la figure 4.

Un partitionnement des vecteurs lors de la quantification, utilisé conjointement à des algorithmes de réduction de dimensionnalité, imposera de recalculer les histogrammes à chaque nouvel ajout ou suppression de mots.

Images entières ou région contenant l'objet. Pour faciliter l'apprentissage, et accélérer le processus de création de vocabulaire, nous avons utilisé les annotations fournies sur les images pour extraire les objets de chaque image selon leur boîte englobante⁶. C'est notre nouvelle base d'apprentissage "découpée".

Nous extrayons une centaine de descripteurs par image, pour un total de 300 000 points.

Les résultats obtenus sont comparés sur la base "découpée", ainsi que sur la base "non-découpée". Dans les deux cas, l'apprentissage du classifieur se fait sur la base d'apprentissage initiale, non sur la base découpée.

Dans les deux cas, les résultats sont comparables, comme le montre la figure 5. Cela signifie que l'algorithme est capable de déterminer quelles informations caractérisent les objets, même si la position des objets dans les images d'apprentissage n'est pas connue.

Nombre de mots du vocabulaire. Le nombre d'éléments du vocabulaire choisi est relativement grand pour toutes les expériences proposées.

Ici, nous cherchons à observer l'influence de ce nombre de mots du vocabulaire sur les résultats de la classification. La figure 6 présente l'évolution des performances pour

⁶des imagerie provenant du fond (sans objet) de même taille sont également extraite de manière à représenter le fond

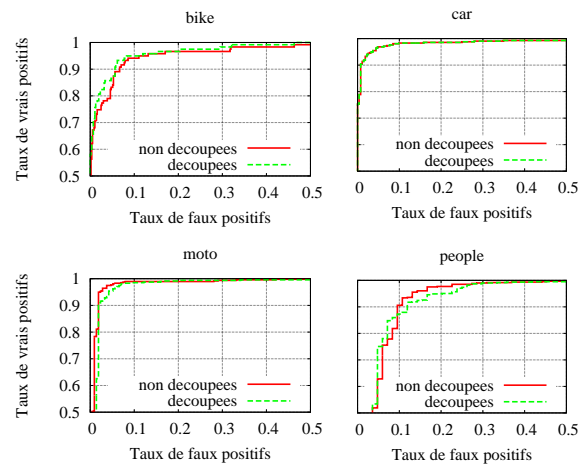


Figure 5: Performances du classifieur selon que les images d'apprentissage sont découpées (pour ne contenir que les objets) ou non

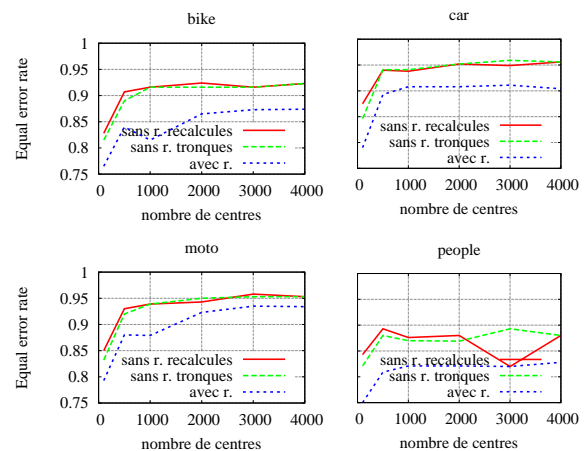


Figure 6: Performances en fonction du nombre de mots produits pour le vocabulaire (voir texte).

3 approches différentes. La première courbe représente les histogrammes recalculés par affectation complète pour le cardinal du vocabulaire choisi. La deuxième courbe représente l'histogramme calculé par affectation totale pour les 4000 éléments du vocabulaire, tout simplement tronqué au nombre d'éléments voulus. Ceci est une approximation discutable, dont on souhaite mesurer les effets. La dernière courbe représente une affectation des points aux centres qui les contiennent dans leur rayon d'influence. Ainsi, seules les colonnes de l'histogramme choisies peuvent être sélectionnées. Cette représentation est exacte sans aucun calcul supplémentaire.

La figure 6 présente les résultats obtenus. Les meilleurs résultats sont bien entendu obtenus pour les histogrammes recalculés, mais l'approximation proposée est acceptable.

Nombre de centres ajoutés par itération. Lors de notre algorithme à deux phases, un certain nombre de centres est choisi lors d'une étape de clustering (voir section 3.2).

L'idéal serait de ne prendre qu'un seul centre à chaque itération, afin d'avoir toujours un échantillonnage optimal, mais ceci est totalement irréalisable, puisqu'il faudrait autant de phases d'échantillonnage, et donc de parcours des données, que d'éléments dans le vocabulaire. Les centres sont donc choisis par groupe dont la taille est un paramètre de l'algorithme.

Nous avons comparé l'évolution des résultats pour des groupes de 10 à 50 vecteurs et n'avons observé aucune diminution des performances, pour cette plage de valeurs.

Nombre de vecteurs échantillonnés par itération. Plus le nombre d'échantillons est grand, plus il est représentatif des données, mais plus la complexité augmente.

Nous avons comparé différentes exécutions de la méthode générative, pour différentes valeurs d'échantillonnage, dans lesquels 20 centres sont extraits. Nous avons fait varier le nombre de vecteurs échantillonnés de 500 à 5000. Nous avons observé que ce paramètre n'a que peu d'influence sur les résultats de la classification, pour les plages de valeur choisies. Le nombre d'échantillons retenus pour le reste des expérimentations est de 3000 par itération.

Comparaison avec une méthode n'utilisant pas d'échantillonnage biaisé. La méthode générative utilisée est comparée à un algorithme comportant également deux phases : l'une d'échantillonnage et l'autre de clustering. La phase de clustering, également réalisée par 'online median', tient compte des centres déjà placés, comme précédemment. La seule différence est que l'échantillonnage est uniforme, et non plus biaisé par un rayon d'influence.

Ces expériences ont permis de mettre en évidence l'importance de l'échantillonnage biaisé qui guide le clustering, et évite de marquer plusieurs fois des régions fortement peuplées. Il dirige les centres vers des régions rares, mais qui s'avèrent informatives. Les résultats sont donc meilleurs avec le rééchantillonnage biaisé.

Comparaison avec une méthode de sélection aléatoire. La méthode avec échantillonnage biaisé est comparée avec une méthode qui choisit de façon complètement aléatoire les éléments du vocabulaire, par tirage parmi tous les descripteurs possibles de la base d'apprentissage.

Comme nous le constatons figure 7, les résultats obtenus avec des centres aléatoires sont moins bons. Notons cependant leur bon niveau de performances.

D'après les expériences menées sur des bases réduites, les méthodes aléatoires possèdent une grande variance entre plusieurs exécutions. Il faudrait répéter cette expérience plusieurs fois pour voir si les performances de la méthode aléatoire décroissent, ou si la base s'avère manipulable par de telles méthodes.

Comparaison avec *k-means*. Nous avons comparé les résultats de notre méthode avec une version "streaming" de *k-means*. Les résultats présentés figure 8 montrent la supériorité de notre méthode.

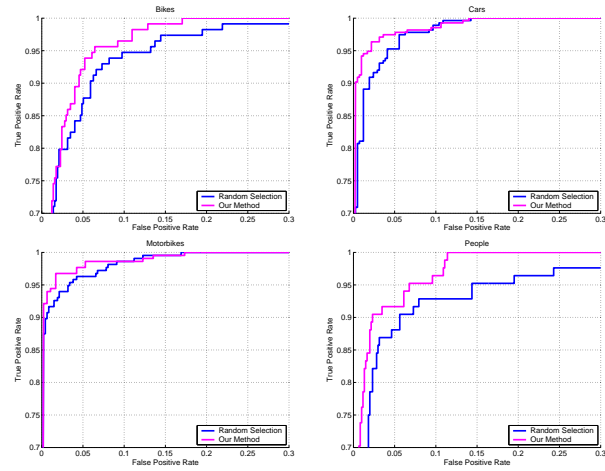


Figure 7: Comparaison avec une sélection aléatoire du vocabulaire

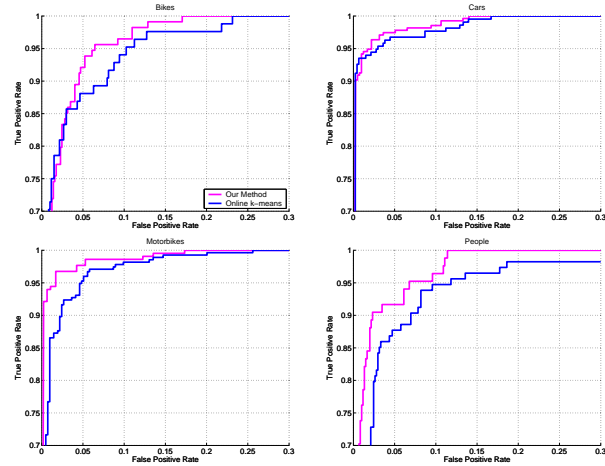


Figure 8: Comparaison avec *k-means*.

Réduction de dimensionnalité et catégorisation d'images. La figure 9 présente les résultats de classification en fonction du nombre d'éléments du vocabulaire, après sélection des mots les plus pertinents.

Cette fois, les meilleurs éléments parmi les 4000 produits sont choisis, à l'aide de 3 méthodes différentes : l'information mutuelle, l'odd-ratio, et une sélection directement faite par le classifieur SVM [2]. Cette dernière méthode donne généralement les meilleurs résultats.

Ces expériences ont été réalisées avec la méthode proposée (rééchantillonnage biaisé), avec utilisation du rayon d'influence pour le calcul des histogrammes. Nous pouvons ainsi extraire les coordonnées de l'histogramme correspondant aux éléments du vocabulaire sélectionnés, pour obtenir les histogrammes du vocabulaire réduit, sans aucun autre calcul.

Il s'avère que très peu de primitives, judicieusement choisies, suffisent à s'approcher des résultats obtenus par le vocabulaire complet.

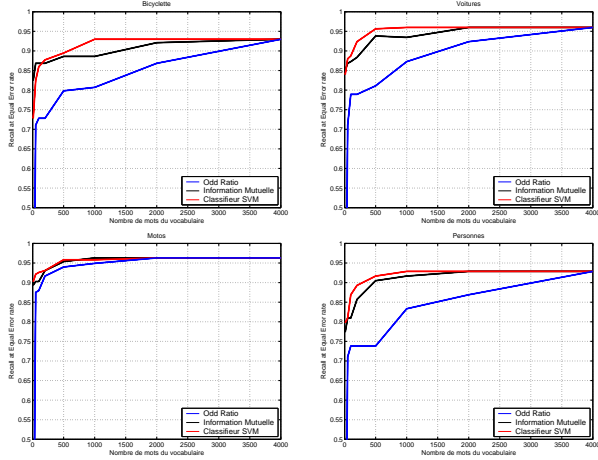


Figure 9: Sélection de primitives par trois méthodes différentes (voir texte).

4.3 Résultats de la compétition Pascal VOC

Huit équipes ont concouru pour la classification, essayant chacune différentes variantes de leur méthode. Seul la meilleure sera présentée. Nos résultats, correspondent à la ligne INRIA 1. INRIA 2 correspond à d'autres participants de notre équipe LEAR (utilisant une méthode proche).

	bike	motorbikes	cars	people
Aachen	0.868	0.925	0.94	0.861
Darmstadt	-	0.644	0.856	-
Edinburgh	0.689	0.793	0.722	0.571
HUT	0.816	0.909	0.921	0.857
INRIA 1	0.93	0.961	0.977	0.917
INRIA 2	0.93	0.937	0.964	0.917
METU	0.781	0.84	0.903	0.803
MPIT	0.754	0.831	0.875	0.731
Southampton	0.895	0.913	0.949	0.881

Nous pouvons constater la supériorité de l'approche que nous proposons par rapport aux meilleures méthodes du moment.

4.4 Expériences complémentaires

Nous avons complété cette étude par des tests réalisés sur la base TUGraz, constituée de 4 classes : les vélos, les voitures, les personnes ainsi que des images de fond. Chacune des classes contient 300 images que nous avons arbitrairement séparées en une base d'apprentissage, utilisée pour la construction du vocabulaire et l'entraînement du classifieur, et une base de test pour l'évaluation, de 150 images chacune.

L'extraction des descripteurs est faite selon une grille régulière en position et en échelle, suivant les mêmes paramètres que précédemment. Le vocabulaire est construit en utilisant les paramètres de l'algorithme donnant de bons résultats pour la base précédente : à savoir 3000 échantillons, un rayon d'influence de 0.6, et 20 centres ajoutés à chaque itération.

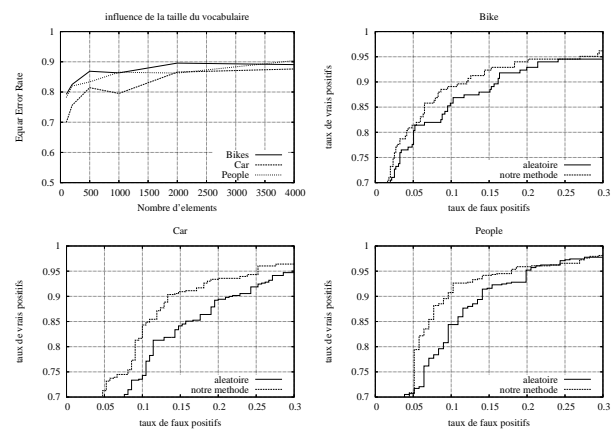


Figure 10: Expériences menées sur TUGraz

Influence de la taille du vocabulaire. Les résultats présentés sur la première courbe, figure 10 confirment qu'une augmentation du vocabulaire améliore les performances, mais de façon relativement peu significative.

Comparaison avec un vocabulaire aléatoire. Les performances de classification sont améliorées par rapport à l'utilisation d'un vocabulaire aléatoire, pour les différentes tailles de vocabulaire testées. Les courbes ROC des résultats pour un vocabulaire de 4000 mots sont présentées dans les 3 dernières courbes de la figure 10.

5 Conclusions et travaux à venir

Nous avons proposé dans cet article une méthode efficace de production de vocabulaires visuels, à partir de descripteurs échantillonnés uniformément dans une base d'images d'apprentissage.

Il s'agit d'une méthode de type génératif, visant à couvrir de manière optimale l'espace des descripteurs extraits des images par le vocabulaire, mais favorisant les zones de faible densité.

Ces différentes idées ont été confrontées à une base d'images significative, créée à l'occasion du *VOC Pascal Challenge*, une compétition en classification et en détection. Les résultats présentés dans cet article montrent la pertinence de l'approche que nous venons de proposer.

Cependant, si ce défi pouvait paraître ambitieux lors de son lancement au printemps dernier, les résultats présentés dans cet article atteignent des performances si hautes qu'il est désormais difficile d'évaluer l'apport de telle ou telle nouvelle idée.

Les techniques de réduction de dimensionalité proposées sont efficaces, et l'utilisation d'une centaine de mots pertinents suffit à obtenir des résultats proches des résultats obtenus avec le vocabulaire complet.

Nous avons pu montrer au cours de cet article le bien fondé de la démarche proposée : le système peut apprendre automatiquement ce qui caractérise visuellement les objets, même sans savoir a priori où ils se trouvent dans les images d'apprentissage. Cela ouvre la porte à des applica-

tions nombreuses.

De telles applications ne pourront voir le jour que si des avancées nouvelles sont faites. Ces avancées constituent des perspectives à plus long terme de notre travail. Comment faire que de tels systèmes apprennent à partir de moins d'images ? Comment discriminer des catégories nombreuses ? Comment modéliser des informations plus complexes que la simple présence d'objets, comme par exemple les relations entre les objets ? Comment combiner les informations visuelles extraites avec des informations textuelles ou numériques liées aux images ? Toutes ces questions, parmi de nombreuses autres, constituent des ouvertures possibles et nécessaires à cette problématique très riche qu'est la catégorisation d'images.

References

- [1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *PAMI*, 26(11):1475–1490, November 2004.
- [2] J. Brank, M. Grobelnik, N. Milic-Frayling, and D. Mladenic. Interaction of feature selection methods and linear classification models. In *Proceedings of the ICML Workshop on Text Learning*, 2002.
- [3] N. Chawla, N. Japkowicz, and A. Kolcz, editors. *Proceedings of the ICML Workshop on Learning from Imbalanced Data Sets*, 2003.
- [4] G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV workshop on Statistical Learning in Computer Vision*, pages 59–74, 2004.
- [5] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV*, pages 634–640, 2003.
- [6] M. Everingham, L.V.Gool, C. Williams, and A. Zisserman. Pascal visual object classes challenge results. <http://www.pascal-network.org/challenges/VOC/voc/>, 2005.
- [7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, pages II: 264–271, 2003.
- [8] D.A. Forsyth and J. Ponce. Computer vision: A modern approach. In *Prentice-Hall*, 2003.
- [9] D. Geman and A. Koloydenko. Invariant statistics and coding of natural microimages. In *IEEE Workshop on Statistical and Computational Theories of Vision*, 1999.
- [10] B. Georgescu, I. Shimshoni, and P. Meer. Mean shift based clustering in high dimensions: A texture classification example. In *ICCV*, pages 456–463, 2003.
- [11] D. Hall and J.L. Crowley. Detection du visage par caractéristiques génériques calculées à partir des images de luminance. In *Reconnaissance des Formes et Intelligence Artificielle*, 2004.
- [12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.
- [13] B. Heisele, P. Ho, J. Wu, and T. Poggio. Face recognition: component-based versus global approaches. *CVIU*, 91(1-2):6–21, July 2003.
- [14] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Tenth European Conference on Machine Learning ECML-98*, pages 137–142, 1999.
- [15] F. Jurie and C. Schmid. Scale-invariant shape features for recognition of object categories. In *CVPR*, pages II: 90–96, 2004.
- [16] F. Jurie and W. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*. 2005.
- [17] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC03*, 2003.
- [18] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, June 2001.
- [19] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
- [20] R. R. Mettu and C. G. Plaxton. The online median problem. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 339, 2000.
- [21] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV*, 2002.
- [22] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. pages 257–263, 2003.
- [23] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.
- [24] M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classification. In *ICCV*, pages 281–288, 2003.
- [25] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV*, pages I: 18–32, 2000.
- [26] Gary M. Weiss. Mining with rarity: a unifying framework. *SIGKDD Explor. Newsl.*, 6(1):7–19, 2004.
- [27] S.C. Zhu, C.E. Guo, Y. Wang, and Z. Xu. What are textons? *IJCV*, 62(1-2):121–143, April 2005.