Méthodes pour l'interprétation automatique d'images en milieu urbain

Nicolas Hascoët

THESE DE DOCTORAT TELECOM SUDPARIS

Spécialité : Informatique et Télécommunications

École Doctorale : Informatique, Télécommunications et Électroniques de Paris

Présentée par

Nicolas HASCOËT

Pour obtenir le grade de
DOCTEUR TELECOM SUDPARIS

METHODES POUR L'INTERPRETATION AUTOMATIQUE D'IMAGES EN MILIEU URBAIN

Devant le jury composé de :

Présidente :	Madame Catherine ACHARD, Maître de conférences, HDR à l'Université Pierre-et-Marie-Curie
Rapporteur :	Monsieur Mohamed DAOUDI, Professeur à Télécom Lille
Rapporteur :	Monsieur Mihai CIUC, Professeur à l'Université Polytechnique de Bucarest
Examinateur :	Monsieur Alain VAUCELLE, Docteur chargé de mission à Plaine Commune
Directeur de thèse :	Monsieur Titus ZAHARIA, Professeur à Télécom SudParis

27 juin 2017

N° de thèse : 2017TELE0004

REMERCIEMENTS

J'aimerais tout d'abord remercier mon directeur de thèse, Monsieur le Professeur Titus ZAHARIA pour son intérêt, son soutien et ses nombreux conseils pendant la durée de ma thèse et notamment sa relecture méticuleuse lors de la rédaction de mon manuscrit final m'aidant sans aucun doute à préciser mon propos.

Je remercie également Messieurs les Professeurs Mohamed DAOUDI et Mihai CIUC pour leur intérêt porté à mes recherches et pour avoir accepté la lourde tâche de rapporter mon travail de thèse.

Mes remerciements vont également à Madame Catherine ACHARD et Monsieur Alain VAUCELLE pour avoir accepté de participer à ce jury de thèse en tant qu'examinateurs.

Madame le Professeur Béatrice PESQUET-POPESCU et Monsieur le Professeur Azeddine BEGHDADI ont accepté d'être présents à ma soutenance de mi-parcours et je les en remercie. Je leur suis aussi reconnaissant pour les conseils qu'ils m'ont apportés à l'issu de cette évaluation.

Ce travail n'aurait pu être mené à bien sans la disponibilité et l'aide apportées par Monsieur Andrei BURSUC, Monsieur Hugo BOUJUT, Madame Ruxandra TAPU et Monsieur Bogdan MOCANU. Leurs nombreux conseils et leurs expertises m'ont permis de mieux appréhender ce domaine et de mettre en œuvre différentes phases d'expérimentation de mes travaux.

Merci aussi à mes collègues pour leur soutien, leur présence et leur bonne humeur quotidienne.

J'adresse aussi mes remerciements à Madame Evelyne TARONI et Madame Véronique GUY pour leur soutien et leur aide entre autre de gestion administrative.

Ce travail n'aurait pas été possible sans le soutien de Télécom SudParis, de l'Institut Mines Télécom et de l'École Doctorale d'Informatique, Télécommunication et Électronique.

Je n'oublie pas Messieurs Michel SIMATIC, Denis CONAN et le Monsieur le Professeur Bruno DEFUDE, ainsi que mon père, le Professeur Jean-Yves HASCOËT pour leurs conseils judicieux avant de commencer et pendant ce travail de recherche.

Je remercie enfin ma famille, mes amis et tout particulièrement ma mère pour leurs encouragements et leur compréhension qui m'ont soutenu tout au long de cette thèse.

TABLE DES MATIERES

LISTE DES FIGURES

LISTES DES TABLEAUX

RESUME

Cette thèse présente une étude pour l'interprétation automatique d'images en milieu urbain. Il s'agit ici de proposer une application permettant de reconnaître différents monuments dans diverses images présentant des scènes complexes. En étudiant les modèles de représentation d'images basés sur des points d'intérêts locaux, la problématique principale est de différencier les informations locales extraites du bâtiment recherché parmi tous les points extraits de l'image. En effet, la particularité d'une image de bâtiment vient de la nature publique de la scène. L'objet que l'on cherche à identifier est au milieu de divers autres objets pouvant interférer avec ce dernier. Nous pouvons citer par exemple, la végétation, des piétons ou des véhicules mais encore d'autres bâtiments environnants ne faisant pas partie du monument recherché.

Nous présentons dans une première partie un état de l'art des méthodes de reconnaissance d'images en se concentrant sur l'utilisation de points d'intérêts locaux ainsi que des bases de données pouvant être employée lors des phases d'expérimentation. Nous retenons au final le modèle de sac de mots (BOW) appliqué aux descripteurs locaux SIFT (S*cale-Invariant Feature Transform*).

Dans un second temps nous proposons une approche de classification des données locales faisant intervenir le modèle de machine à vecteurs de support (SVM). L'enjeu est ici d'isoler les points extraits du bâtiment d'intérêt recherché dans une image et d'en rejeter ceux extraits du reste de l'image. La première méthode exposée classifie l'information locale des divers points extraits d'une image grâce à un entraînement faisant intervenir un nombre très limité de données (dans la phase d'expérimentation, pour une base de données de 6 000 images, seules 55 sont utilisées en entraînement). Sachant que différentes catégories de bâtiments sont présentes dans une même base de données, différentes stratégies d'entraînement à la classification sont développées. Le premier modèle ne présente pas de distinction entre les différentes catégories de bâtiments dans la base de données et classifie chaque monument par rapport aux autres objets d'une image. Le deuxième modèle exposé traite chaque catégorie distinctement et calcule donc un modèle de classification différent pour chaque catégorie. La problématique est alors de définir le critère de sélection du modèle adéquat pour une image donnée. Les deux critères retenus sont le nombre total de points prédits d'une part et le critère de confiance de la prédiction.

Une troisième partie suggère l'ajout d'une correction géométrique de la classification obtenue précédemment. Nous obtenons ainsi une classification non seulement de l'information locale mais aussi visuelle permettant ainsi une cohérence géométrique de la distribution des points d'intérêt. En effet, l'étape précédente ne prend en compte que l'information locale sans soucis du voisinage. Ainsi, l'influence de la prédiction des voisins d'un point peut permettre d'éviter des points d'intérêts isolés.

Enfin, un dernier chapitre présente les résultats expérimentaux obtenus pour ces différentes méthodes en comparaison avec les méthodes de l'état de l'art. Nous sélectionnons les bases de données des bâtiments de Paris et d'Oxford comptant respectivement 6412 et 5062 images pour 11 catégories de monuments différents. Les méthodes de représentation d'images expérimentées sont le modèle de sac de mots et le descripteur global VLAD (*Vector of Locally Aggregated Descriptors*).

ABSTRACT

This thesis presents a study for an automatic interpretation of urban images. We propose an application for the retrieval of different landmarks in images representing complex scenes. The main issue here is to differentiate the local information extracted from the key-points of the desired building from all the points extracted within the entire image. Indeed, an urban area image is specific by the public nature of the scene depicted. The object sought to be identified is fused within various other objects that can interfere.

First of all, we present a state of the art about image recognition and retrieval methods focusing on local points of interest. Databases that can be used during the phases of experimentation are also exposed in a second chapter. We finally retain the Bag of Words modèle applied to local SIFT descriptors.

In a second part, we propose a local data classification approach involving the Support Vector Machine model. The interest shown with this proposed approach is the low number of data required during the training phase of the models. Different training and classification strategies are also discussed.

A third step suggests the addition of a geometric correction on the classification obtained previously. We thus obtain a classification not only for the local information but also for the visual information allowing thereby a geometric consistency of the points of interest.

Finally, a last chapter presents the experimental results obtained, in particular involving images of buildings in Paris and Oxford.

I. INTRODUCTION

I.1. Contexte

La vision par ordinateur a pour but de permettre à un terminal (par exemple un ordinateur ou un *smartphone* aujourd'hui) d'analyser, de traiter et de tenter de « comprendre » une ou plusieurs images (voir des vidéos) grâce un système d'acquisition (composé souvent d'une ou de plusieurs caméras). Il s'agit de tenter d'imiter la vision humaine, tout en exploitant les améliorations techniques qui ont émergé les dernières années : champ de vision large, vision nocturne, acquisition de cartes de profondeur, hautes résolutions…

Les applications sont diverses tant dans le domaine industriel (exemples : contrôle de qualité ou automatisation des moyens de production) que dans le domaine de la recherche pour permettre aux ordinateurs, smartphones et robots de « voir » et comprendre le monde qui les entoure.

Nous nous intéressons ici plus particulièrement à l'interprétation des scènes urbaines, l'enjeu étant de reconnaître les bâtiments présents dans une scène urbaine acquise avec une caméra arbitraire.

Sur la même lignée de technologies émergentes liées à la vision par ordinateur, citons, à titre d'exemple type, les nombreuses applications qui existent aujourd'hui dans le domaine de la reconnaissance faciale. On en retrouve particulièrement embarquées dans les *smartphones* tant pour améliorer les prises de photos et notamment de *selfies* que pour renforcer la sécurité avec des verrous basés sur la reconnaissance du visage du propriétaire. La reconnaissance faciale est aussi de plus en plus utilisée dans le domaine de la robotique pour favoriser des interactions plus humaines entre robots et individus.

Le domaine de la reconnaissance de monuments et de bâtiments en milieux urbains connaît également, ces dernières années, un développement important. Toutefois, par rapport aux méthodes de reconnaissance de visages, plusieurs difficultés nécessitent d'être prises en compte et traitées. En premier lieu, il s'agit notamment de la grande diversité de bâtiments que l'on souhaite reconnaître, de formes, textures et complexités différentes. Ainsi, le nombre de points remarquables que l'on puisse détecter sur un bâtiment est beaucoup plus important que sur un visage. En outre, les tailles de bases de données de référence sont bien plus importantes. De plus, en raison de l'éventuelle répétition de motifs sur un ou plusieurs bâtiments, les points d'intérêt peuvent devenir moins discriminants.

La reconnaissance de bâtiments trouve son application dans différents domaines comme la géolocalisation, la cartographie, pour les voitures autonomes ou encore les moteurs de recherche. En effet, quand les technologies basées sur la technologie GPS ne fonctionnent pas (à cause d'interférences, de manque de signal, etc.), reconnaître visuellement un bâtiment, un monument ou un quartier permet de se géo-localiser par rapport à ce bâtiment reconnu.

Il en va de même dans le domaine de la cartographie où le but est de distinguer les bâtiments les uns des autres et de leur environnement, en s'appuyant sur des technologies de type Système d'Information Géographique (SIG).

Dans le but de permettre à un automate d'analyser et de donner du sens à son environnement via un système d'acquisition visuel, la problématique de reconnaissance de bâtiments dans un environnement urbain y trouve une place centrale. Un exemple aujourd'hui sont les voitures autonomes comme la *Google Car*, la voiture sans conducteur de *Google*. En effet, grâce à plusieurs caméras, cette voiture doit pouvoir acquérir, analyser et interpréter son environnement. Il s'agit bien ici d'un environnement public, urbain et complexe. Dans ce cadre, il est impératif que la voiture puisse reconnaître les bâtiments, les piétons, les autres voitures, la route et ce de façon automatique et au plus proche du temps réel.

Enfin, dans une utilisation quotidienne, les moteurs de recherche sont des outils nécessaires au traitement d'un grand nombre de données. Aujourd'hui, le problème n'est plus d'avoir accès à une information mais de trouver l'information adéquate parmi l'immense quantité de données. Bien que les moteurs de recherche utilisant des requêtes à base de texte (comme le célèbre moteur de recherche de *Google*) sont maintenant capables de retrouver des documents correspondants à une requête de mots clefs de façon suffisamment précise et exacte, le domaine de récupération de données en arrive maintenant à la problématique d'utiliser des images en tant que requête.

Nous nous intéresserons ici à cette dernière problématique d'utilisation d'images dans un cadre de moteur de recherche pour des images acquises en milieu urbain.

I.2. Objectifs

Différents modèles et méthodes sont aujourd'hui disponibles dans le domaine de la recherche de documents à partir de requêtes textuelles par mots clefs. A partir de ces modèles de représentation de grandes bases de données de documents de nouvelles approches ont été conçues de façon analogue pour le domaine de l'imagerie.

L'objectif de nos travaux est de proposer un modèle de représentation d'image adapté aux environnements urbains et capable d'améliorer le processus de recherche automatique de bâtiments à partir d'une image requête. Nous plaçons cette problématique dans un cadre urbain *i.e.* dans lequel les images prises en compte sont acquises dans un environnement extérieur et publique pour la reconnaissance de bâtiments et monuments touristiques.

Les images considérées sont prises en extérieur, dans des conditions d'acquisition complètement libres. Cela pose la question de l'influence de l'éclairage et de la météo sur la persistance et la stabilité de l'information présente sur une image. En effet, la Tour Eiffel prise en photo de jour ou de nuit, par beau temps ou par temps de pluie ne semble pas être la même visuellement et pourtant nous reconnaissons bien un seul et même monument. Un enjeu de taille pour la reconnaissance automatique d'images est donc de vérifier si cette persistance d'interprétation du contenu peut être intégrée dans les modèles mathématiques de représentation.

De plus, s'agissant de bâtiments, une photo peut être prise à partir de différents points de vue et surtout de plus ou moins loin du monument en question. Dans ce cadre, il est donc indispensable de considérer et de traiter les aspects d'invariance par rapport à la *pose* (*i.e.*, la position et l'orientation relative par rapport à l'observateur) de l'objet d'intérêt. Par exemple, une photo de la Tour Eiffel peut être acquise en contre plongée de très près ou de très loin par une vue panoramique. Une image peut la représenter d'un côté ou d'un autre. Dans ce cas, de même que précédemment, l'objet que nous

recherchons (la Tour Eiffel dans cet exemple) n'est visuellement pas le même alors qu'il s'agit en réalité du même bâtiment (**Figure 1**).

Figure 1. Exemples d'images de différentes poses de la Tour Eiffel.

Le problème est le même que dans le cas des conditions différentes d'éclairage.

D'autre part, nous nous intéressons ici aux images acquises dans un environnement non seulement urbain mais public. Dans ce cadre, une image n'inclut pas seulement l'objet de notre requête mais présente également d'autres éléments/objets qui rajoutent ainsi de l'information pouvant aider ou au contraire nuire à la reconnaissance de l'objet d'intérêt. Par exemple, si l'objectif est de retrouver les photos de la Tour Eiffel, celles-ci ne montrent pas uniquement le monument mais incluent aussi des éléments perturbateurs, qui peuvent être les bâtiments environnants, la végétation aux alentours, les piétons et les véhicules aux abords… (**Figure 2**).

Figure 2. Exemples d'autres objets visibles dans une scène urbaine complexe représentant la Tour Eiffel.

La problématique est donc ici d'isoler le bâtiment recherché afin de préciser la requête, la rendre plus pertinente et ne pas surcharger inutilement les informations extraites des images.

Ces différents éléments de discussion montrent la complexité des scènes urbaines pour la reconnaissance d'images de façon automatique et l'enjeu de pouvoir améliorer et optimiser ce processus.

I.3. Applications

Avec les moteurs de recherche grands publics existants, nous pouvons accéder aujourd'hui aux documents correspondant à des requêtes textuelles par mots clefs. Diverses méthodes et algorithmes ont été mis en place pour relier des mots à des documents qui sont par la suite retournés en fonction de mots requête et classés par la suite par ordre de pertinence. Toutefois, cette problématique de récupération de documents au sens large est maintenant un défi croissant dans le domaine de l'imagerie.

Dans ce cadre, l'objectif est de récupérer les images les plus proches à une requête fournie comme exemple. De cette façon, le problème est de trouver les "*bons" mots visuels* pour décrire une image. L'objectif est d'élaborer de modèles de représentation d'image, de façon à pouvoir analyser et traiter une image d'une façon similaire à la vision humaine. Cela nécessite à la fois la mise en œuvre de méthodes de vision par ordinateur, pour déterminer les descripteurs d'image pertinents, et d'apprentissage statistique, pour simuler la capacité du cerveau humain à inférer de l'information et de la connaissance à partir d'un ensemble de données incomplètes.

Ces processus requièrent également la disponibilité de bases de données adéquates. Dans notre cas, nous utilisons des bases de données de référence regroupant des monuments touristiques de différentes villes (*cf.* **Chapitre III**).

De multiples approches ont été proposées dans l'état de l'art pour isoler les points clés significatifs qui puisse décrire à la fois localement et globalement une image, d'une façon discriminante. L'objectif n'est pas seulement de détecter des formes d'objets dans une image, mais de donner un sens sémantiquement pertinent pouvant conduire à une interprétation automatique d'une scène donnée. Parmi les approches les plus populaires dédiées à ces objectifs citons les techniques par descripteurs locaux comme les *Scale Invariant Feature Transform* (SIFT) [Lowe04] et des représentations d'images globales telles que le modèle de sacs de mots (en anglais *Bag of Words* ou BOW) [SiZi03] ou encore le modèle de *Vectors of Locally Aggregated Descriptors* (VLAD) [JDSP10]. Dans ce cadre, toutes les images peuvent être décrites avec une information formatée pour différentes applications, incluant comparaison d'images et recherche d'objets dans une image.

Dans le contexte d'application de recherche d'objets, le défi est de reconnaître un objet spécifique présent dans une image et pouvant faire apparaître un environnement relativement complexe (**Figure 3**).

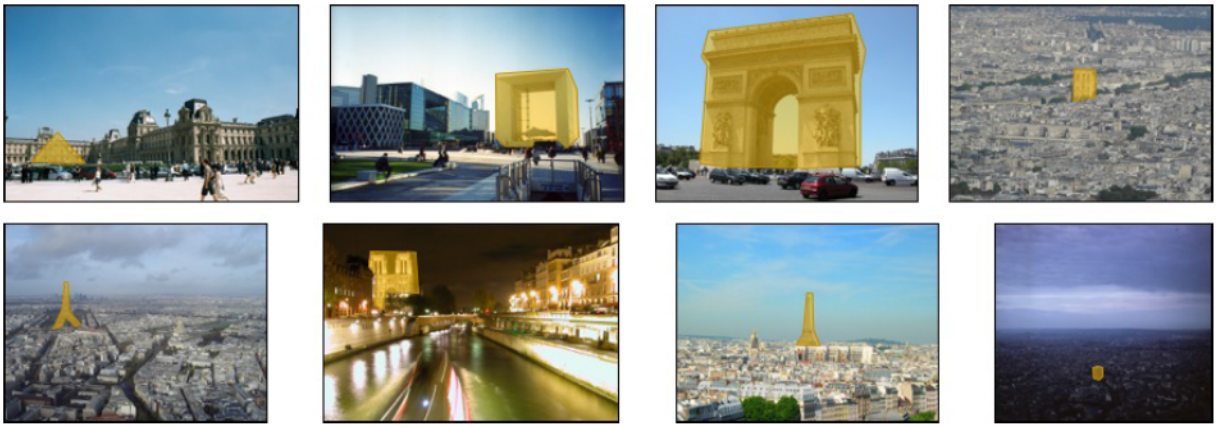


Figure 3. Exemples d'images de bâtiments d'intérêts de la base de données Paris6k [PCIS08] au milieu d'autres objets dans différentes scènes urbaines complexes.

La reconnaissance de bâtiments/monuments que nous traitons dans cette thèse s'inscrit dans le cadre de cette problématique de reconnaissance d'objets. Les difficultés spécifiques à lever concernent la complexité des images. Ainsi, les photos sont prises en plein air, dans des espaces publics, où d'autres objets peuvent interférer avec le bâtiment recherché en requête. Les piétons, les véhicules, la végétation et d'autres bâtiments environnants peuvent aussi occulter l'objet d'intérêt dans l'image ou même ajouter des données indésirables pour la reconnaissance de cet objet. De plus, comme nous avons affaire à des monuments touristiques, le point de vue d'où un bâtiment a été pris peut varier significativement d'une image à une autre et, par conséquent, la *pose* de l'objet interrogé peut biaiser l'information locale.

Nous proposons une approche pour aider la reconnaissance et la récupération des images représentant à partir de bâtiments et monuments d'intérêt dans un environnement urbain. Le principe consiste à filtrer les points d'intérêt locaux pour ne garder que les points clés décrivant spécifiquement l'objet d'intérêt. Cette étape est réalisée avant la représentation globale de l'image et mise en œuvre à l'aide de modèles de type BOW et VLAD, de façon à ce que ces représentations globales soient utilisées pour décrire l'objet réel et non l'ensemble de l'image.

Reconnaître un bâtiment dans une image permet de rechercher les images similaires dans une base de données. Une autre application que nous avons retenue concerne la classification automatique d'images. En effet, s'il est possible d'isoler et d'identifier un bâtiment particulier sur chaque image de la base de donnée considérée il devient également possible de regrouper les images représentant le même bâtiment dans une même classe et ceci de façon automatique.

Dans la suite, nous dressons dans un premier temps un état de l'art des techniques de recherche d'image en milieu urbain (**Chapitre II**).

Nous présenterons ensuite les bases de données d'images de référence que nous utilisons dans nos expérimentations pour tester et évaluer de manière objective les approches proposées (**Chapitre III**).

Les trois chapitres suivants introduisent une nouvelle approche de la recherche d'image dont le principe de base est de classifier les points d'intérêt d'une image pour n'en garder que les plus pertinents lors de la phase de recherche. Une première partie (**Chapitre IV**) introduit un modèle global de classification des descripteurs qui s'appuie sur des machines à vecteurs de support (*Support*

Vector Machine - SVM) [BoGV92]. Ensuite, dans le **Chapitre** V nous montrerons que cette catégorisation peut être affinée en faisant intervenir de multiples classifieurs combinés à différentes stratégies de fusion. Finalement, nous introduisons une approche de correction géométrique (**Chapitre VI**) de cette catégorisation qui a comme objectif d'améliorer la cohérence des objets isolés dans une même image.

Enfin, le **Chapitre VII** présente les résultats expérimentaux obtenus avec les différentes approches proposées, comparés aux résultats des méthodes de l'état de l'art, de façon à démontrer et quantifier les améliorations apportées par notre modèle.

II. ÉTAT DE L'ART

Résumé. Ce chapitre dresse un état de l'art dans le domaine de la reconnaissance et de la recherche de bâtiments à partir de scènes urbaines complexes. Il est structuré en trois parties.
Dans un premier temps, nous exposons les principaux modèles de représentation d'images aujourd'hui disponibles et largement utilisés d'une manière plus générale à des objectifs de reconnaissance d'objets. Parmi les modèles présentés, nous nous concentrons notamment sur les représentations par points d'intérêt, avec procédures d'extraction et descripteurs associés (SIFT, SURF…). Les principales méthodes d'agrégation de l'information locale, incluant de modèles de type *Bag of Visual Words* (BOW), vecteurs de Fischer ou représentation VLAD sont également rappelées.
Dans un deuxième temps, nous analysons les principales familles d'approches qui s'attaquent notamment à la problématique complexe et difficile de reconnaissance de bâtiments.
Enfin, la troisième partie s'intéresse aux pistes récentes d'évolution et d'amélioration, qui reposent sur l'exploitation d'information textuelle additionnelles, l'extension de requêtes, la mise en œuvre d'approches par régions d'intérêt, les techniques d'affinage et de réduction des points d'intérêt utilisés, ou encore les méthodes émergentes par réseaux neuronaux profonds.

Mots clés : Points d'intérêt locaux, filtres Gaussiens, gradients d'image, descripteurs locaux, SIFT, SURF, HOG, GIST, classification, BOW, VLAD

Afin de pouvoir reconnaître un bâtiment à partir d'une photo, une étape d'analyse d'image, qui vise à déterminer les éléments porteurs d'information pertinente, est nécessaire. Rappelons en premier lieu les différentes étapes de détection de points d'intérêts, d'extraction et de description de cette information selon différents descripteurs, de regroupement des descripteurs dans un vocabulaire et enfin de requête et de recherche avec la méthode de *Bag of Words* (BOW) [SiZi03].

Des applications de reconnaissance de bâtiments existent aujourd'hui dans différents domaines comme par exemple la cartographie [BiMS09] et [ZhTM13]. D'autres applications et notamment celles dédiées aux voitures autonomes s'intéressent aussi à l'interprétation de la ville (et donc à la reconnaissance de bâtiments) comme la Google Car.

Le domaine de la reconnaissance d'images appliquée à un environnement urbain, c'est-à-dire pour de la reconnaissance de bâtiments est aussi décrit et étudié dans [DSGS12], [BAGN13], [ToFW08], [SuFJ05] et [BiMS09] ainsi que par Google de façon plus générale avec ses *Google Goggles* [Diaz08] et [JaRa15].

Nous présentons dans ce chapitre les différentes approches existantes pour la recherche et la reconnaissance d'images dans un environnement urbain.

II.1. Modèle de représentation d'image

Citons tout d'abord les travaux de [BuZa13a] comme méthode de base pour la recherche d'images. Cette approche s'appuie sur une représentation d'image exploitant le modèle BOW [SiZi03]. La détection et l'extraction des points d'intérêts est réalisée à l'aide de la méthode de détection de régions hessienne-affines covariantes [MiSc04]. Ces points clefs sont ensuite décrits par une information de locale de l'image qui sont ici les descripteurs SIFT [Lowe04]. Cette information locale est ensuite regroupée pour former un vocabulaire de mots visuels utilisé dans le calcul du modèle de BOW. Une image est ainsi représentée par un histogramme de fréquences relatives des mots du vocabulaire. Ce processus d'analyse est illustrée **Figure 4**. Enfin, pour une image requête, les images les plus proches dans la base de données sont retrouvées en comparant les histogrammes calculés à l'étape précédente et classés en fonction de leur similarité avec l'image d'origine, comme illustré **Figure 5**.

Dans cette optique de représentation et de recherche d'image, nous présenterons dans un premier temps les différentes méthodes de détection des points clés. Nous établirons ensuite une liste de descripteurs locaux usuels dans le domaine de la description d'images. Enfin, nous exposerons deux modèles de représentations d'image que sont les modèles de BOW [SiZi03] et de VLAD (*Vector of Locally Aggregated Descriptors*) [JDSP10].

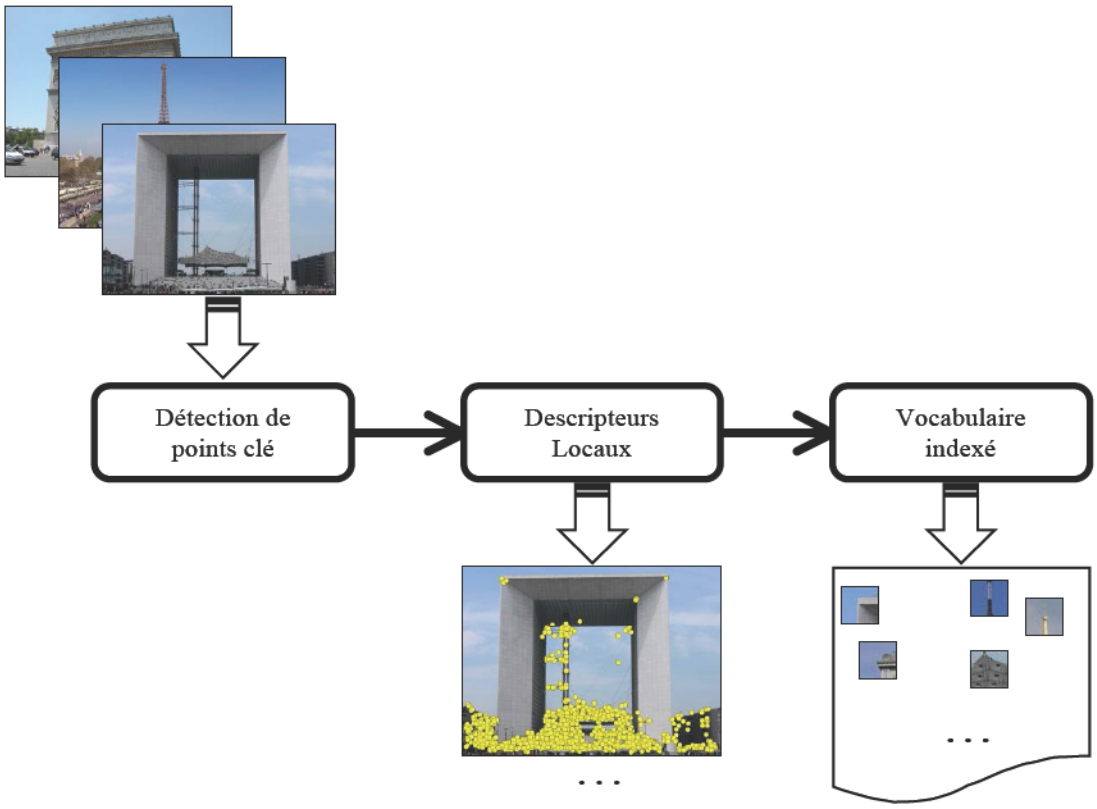


Figure 4. Processus de construction d'un vocabulaire de mots visuels sur l'ensemble des images d'une base de données.

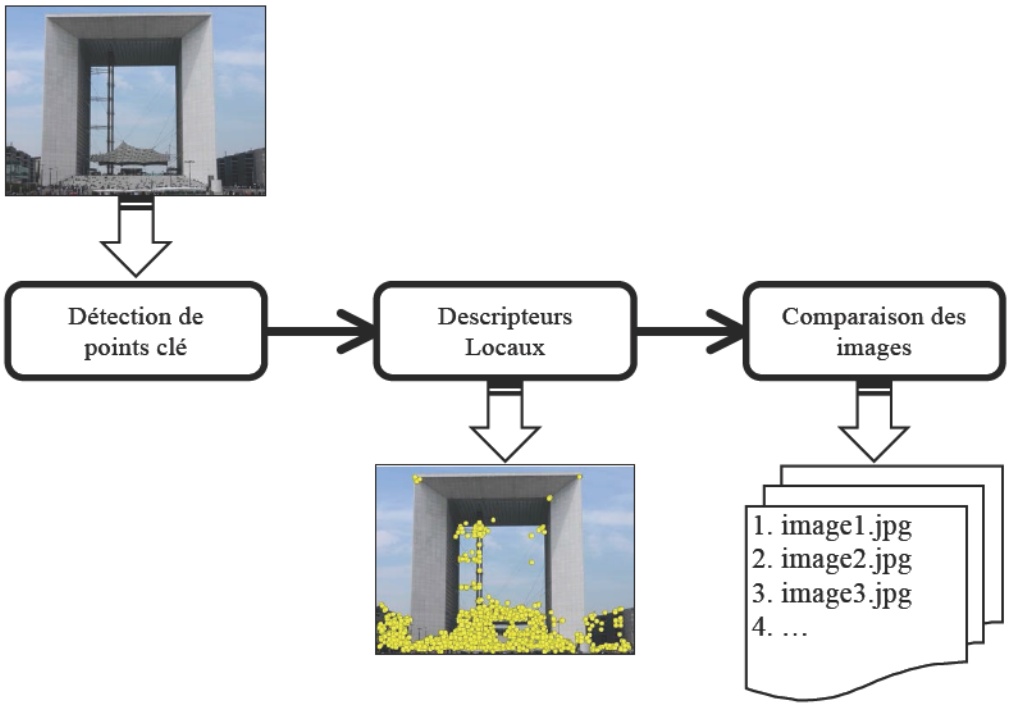


Figure 5. Processus de recherche d'image sur une image requête par comparaisons des informations extraites avec le vocabulaire construit précédemment sur la base de données.

II.1.1. Détection des points d'intérêt

Les points d'intérêt correspondent à une analyse des propriétés locales de l'image. Ainsi, pour la détection de formes dans une image le voisinage d'un coin ou d'un bord d'un objet est *a priori* bien plus riche en termes d'information discriminante qu'un point tiré au hasard au milieu d'une texture. Il s'agit de la première étape dans notre processus de reconnaissance d'images tant dans la partie de construction du vocabulaire que pour la recherche d'images, comme illustré **Figure 6**.

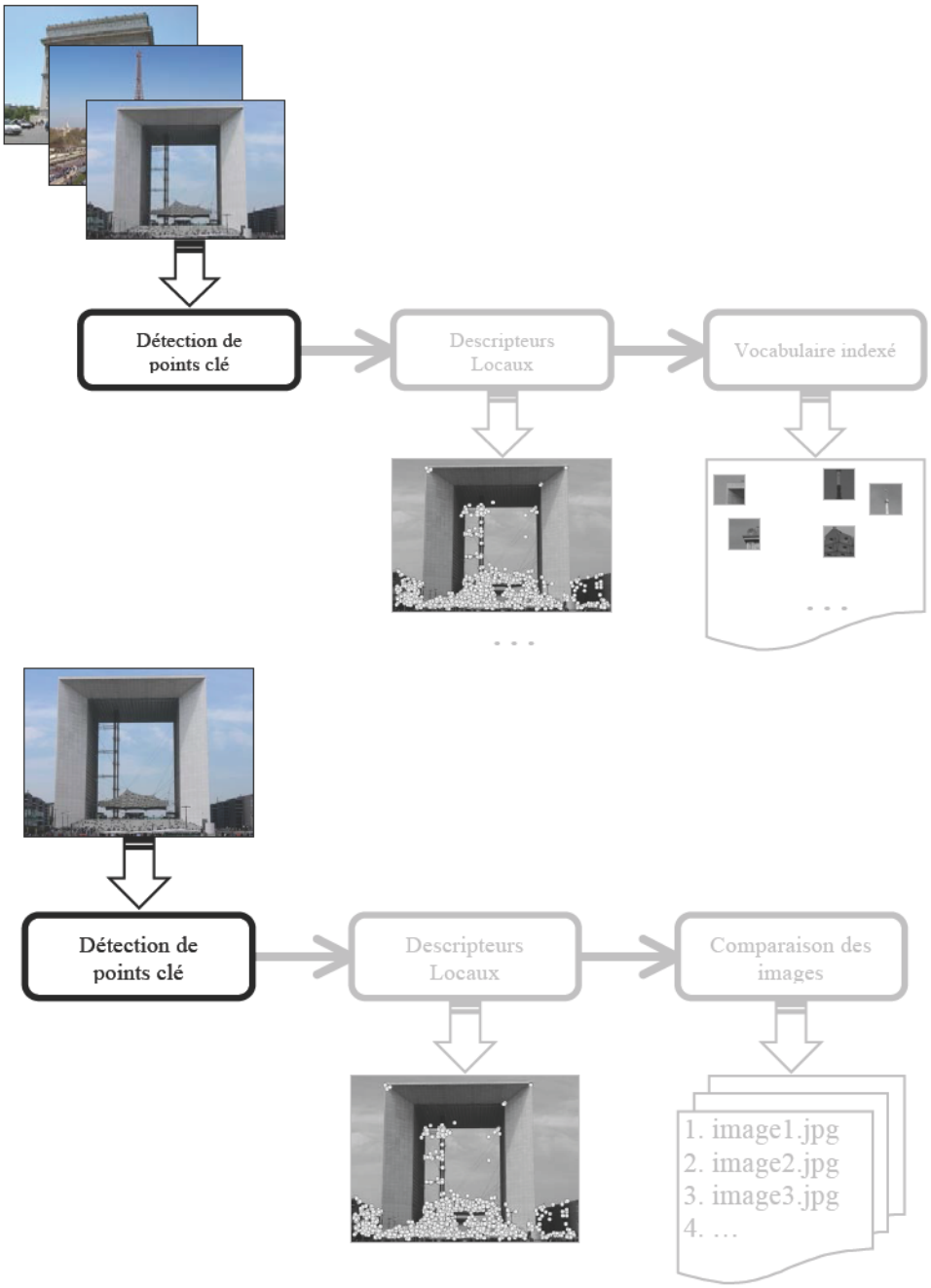


Figure 6. Étapes de détection des points clés dans les processus de construction de vocabulaire et de recherche d'images.

Plusieurs techniques de détection de points clés sont aujourd'hui disponibles. L'objectif est d'assurer une détection fiable et surtout reproductible de points clefs, pouvant supporter une invariance par rapport aux conditions d'acquisitions de l'image et par rapport aux transformations géométriques usuelles que peuvent subir les images (échelle/taille, transformations affine ou perspectives planaires dues aux changements de points de vue/poses…).

II.1.1.1. Détection par filtrage Gaussien

Dans [Lowe04], [MiSc05] et [WiBr07], les auteurs proposent une méthode de détection de points d'intérêts robustes aux changements d'échelle. L'objectif de la détection par filtrage Gaussien est de détecter non pas des points mais les contours des objets (**Figure 7**) qui correspondent à un changement d'intensité lumineuse. Dans cette approche, la méthode de différence de Gaussiennes (ou de Laplacien-Gaussienne) est utilisée pour détecter les formes sur une image. Le principe consiste à comparer une image floutée par un filtre gaussien à sa version originale afin de supprimer les hautes fréquences spatiales, comme illustré **Figure 7**. Nous rappelons l'équation de la fonction gaussienne dans l'équation (1). Par la suite, le filtrage est réalisé selon une convolution (2) de l'image filtrée et sa version originale.

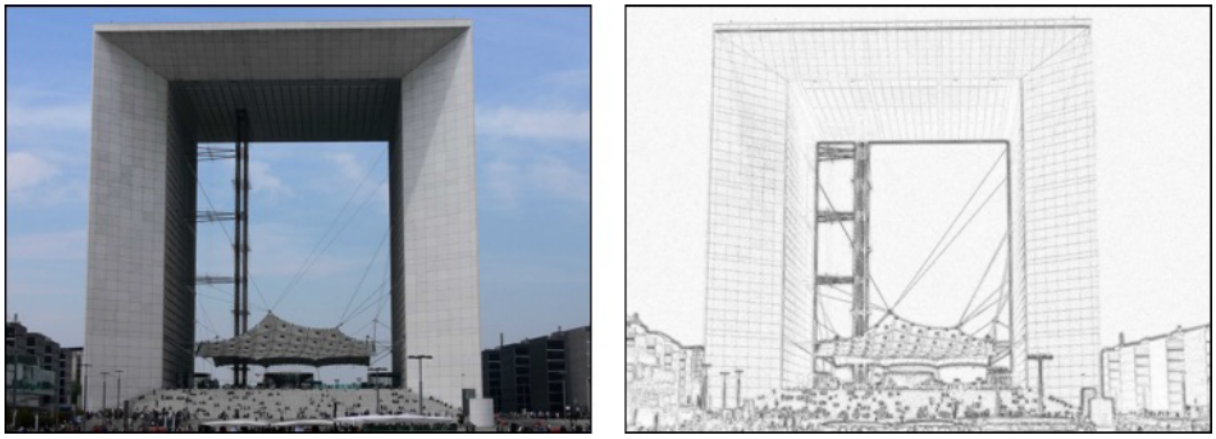


Figure 7. Exemple d'image de l'arche de la Défense et la détection des contours avec la méthode des différences de Gaussiennes.

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (1)$$

$$f * g(x) = \int_{-\infty}^{+\infty} f(t)g(x-t)dt \quad (2)$$

Dans l'équation (1), G correspond au filtre gaussien appliqué au point (x, y) avec σ le paramètre de dispersion du filtre à définir. L'équation (2) rappelle le produit de convolution de deux images f et g au point x.

En variant successivement la variance σ du filtre, il est possible de construire un espace échelle Gaussien qui est notamment utilisé par le détecteur de régions Hessien-affines, décrit dans la section suivante, pour assurer une détection des points clefs à travers multiples résolutions d'images et assurer ainsi à la fois invariance aux changements d'échelle et robustesse au bruit.

II.1.1.2. Détection des points à forte variation de gradients

Plutôt que de détecter tous les contours de formes, le détecteur de Harris [WiBr07] permet de détecter les coins uniquement qui apportent une information structurale et intrinsèquement 2D. Un exemple de détection des coins est illustré **Figure 8**. Ce détecteur est invariant aux rotations mais reste sensible aux changements d'échelle.

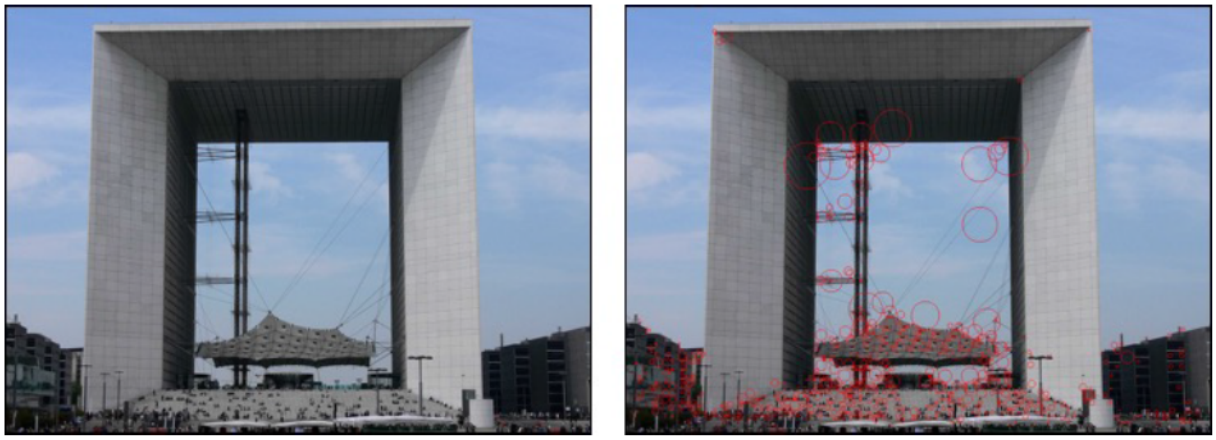


Figure 8. Exemple d'image de l'arche de la Défense et la détection des coins avec la méthode des Hessien-affines.

Les points détectés sont représentés par des ellipses suivant l'orientation du gradient d'image au voisinage du point. Les détecteurs Harris-affine et Hessien-affine ajoutent une étape de normalisation et ces points sont alors représentés par des cercles. Cela permet à ces détecteurs d'être robustes aux transformations affines (rotations, homothéties, similitudes, …). Ils sont présentés et comparés dans [MiSc02], [MiSc04], [KaZB04] et [MTSZ05].

En raison de ses bonnes propriétés à la fois en termes d'invariance et de reproductibilité, le détecteur Hessien-affine présente les caractéristiques les plus adaptées pour la reconnaissance d'images.

En s'appuyant sur le travail des auteurs dans [BuZa13a], dans nos travaux nous avons adopté le détecteur de régions Hessien-affines proposé dans [CDFW04]. Ce détecteur permet d'extraire des points clés sur une image, tout en assurant une robustesse aux changements d'échelle et aux transformations affines.

L'image est traitée dans un espace d'échelle gaussien afin de sélectionner les points invariants aux changements d'échelle et aux transformations affines. Pour une image donnée $I(x, y)$, l'espace d'échelle gaussien associé est la famille des convolutions de $I(x, y)$ avec les filtres gaussiens pour différents seuils donnés. Pour chaque point clé (x, y) sélectionné dans une image I et pour une échelle σ, la matrice Hessienne H est calculée (3). Au final, les points d'intérêt sont les points définissant les extrema locaux du déterminant (4) et de la trace (5) de la matrice Hessienne.

$$H_\sigma(x,y) = \begin{bmatrix} \frac{\partial I_\sigma(x,y)}{\partial x^2} & \frac{\partial I_\sigma(x,y)}{\partial x \partial y} \\ \frac{\partial I_\sigma(x,y)}{\partial y \partial x} & \frac{\partial I_\sigma(x,y)}{\partial y^2} \end{bmatrix} \quad (3)$$

$$Det = \sigma^2\left(H_{xx}H_{yy} - H_{xy}H_{yx}\right) \quad (4)$$

$$Tr = \sigma(H_{xx} + H_{yy}) \quad (5)$$

Il a été montré dans [MTSZ05] que le détecteur de régions hessienne-affines est particulièrement adapté pour des scènes structurées et en particulier pour des images représentant des bâtiments. De plus, comme les photos ont été acquises de divers points de vue, le bâtiment d'intérêt dans la scène peut être présenté sous différentes échelles et différentes poses. Ainsi, le détecteur de régions Hessien-affines semble le plus adapté dans notre contexte de recherche d'image pour la reconnaissance de bâtiments.

Une fois les points d'intérêts détectés, leur voisinage est décrit et représenté à l'aide de descripteurs d'image dédiés, comme décrit dans la section suivante.

II.1.2. Descripteurs locaux

Chacun des points d'intérêts détectés est caractérisé par un vecteur *descripteur*, qui vise à caractériser l'information locale de l'image autour du point d'intérêt considéré. Ceci correspond à la seconde étape dans les processus à la fois de construction de vocabulaire et de recherche d'images (**Figure 9**).

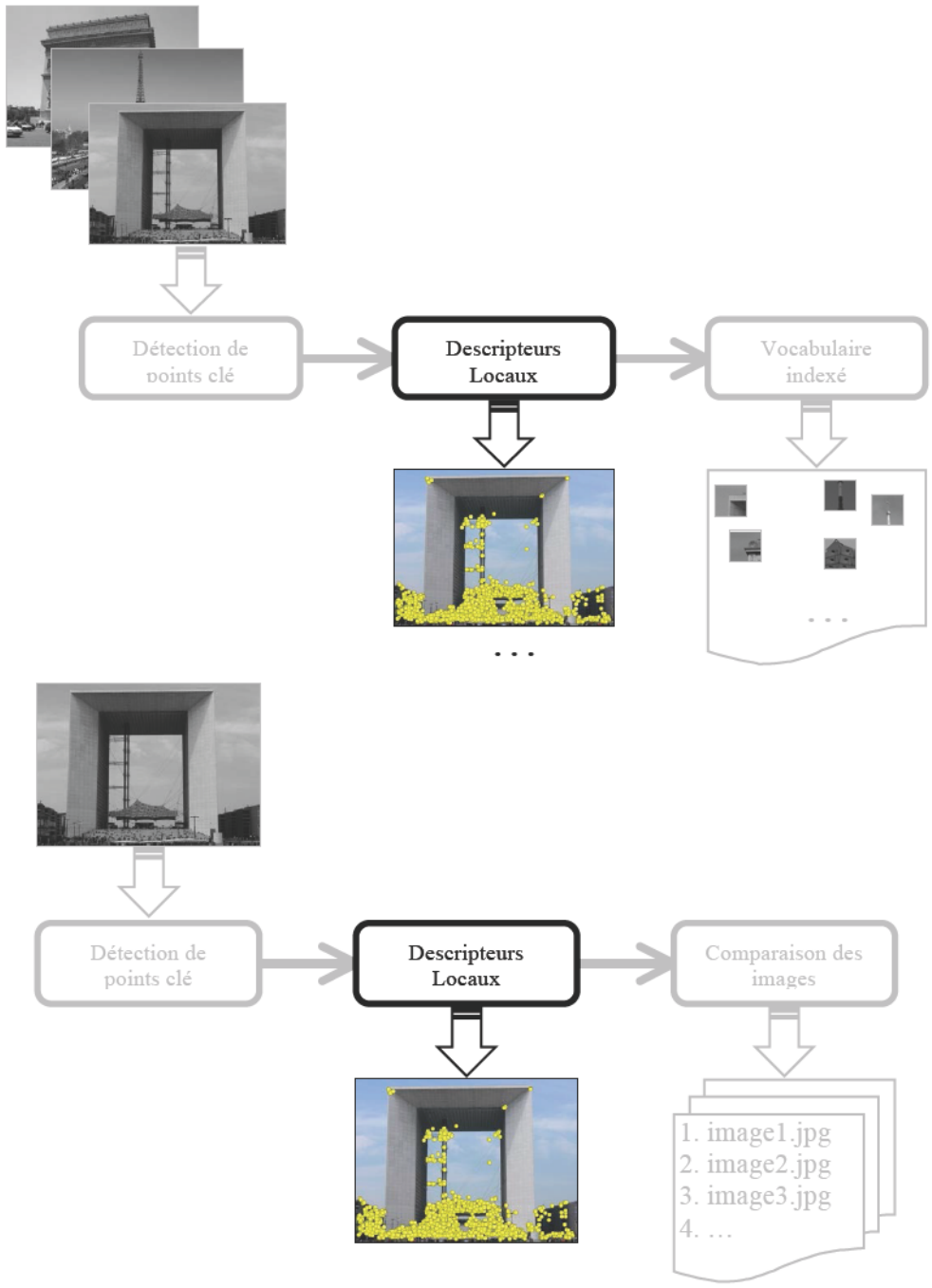


Figure 9. Étapes d'extraction des descripteurs locaux dans les processus de construction de vocabulaire et de recherche d'images.

Parmi les différents descripteurs que l'on peut associer aux points d'intérêt, citons tout d'abord les populaires vecteurs SIFT, rappelés dans le paragraphe suivant.

II.1.2.1. Descripteur SIFT

Les descripteurs SIFT (*Scale Invariant Feature Transform*) [Lowe04] représentent souvent un choix classique dans différents algorithmes de représentation d'image. A chaque point clé détecté

(x, y) est associée à son image gradient L de magnitude m et l'orientation θ comme décrit respectivement dans les équations (6) et (7).

$$m(x,y) = \sqrt{\Delta_x^2 + \Delta_y^2} \ , \tag{6}$$

$$\theta(x,y) = \tan^{-1}\left(\Delta_y/\Delta_x\right) , \tag{7}$$

avec $\Delta_x = L(x+1,y) - L(x-1,y)$ et $\Delta_y = L(x,y+1) - L(x,y-1)$.

Une région de (16 × 16) pixels centrée sur le point clé considéré est définie et ensuite subdivisée en quatre sous-blocs de (4 × 4) pixels, comme illustré **Figure 10**.

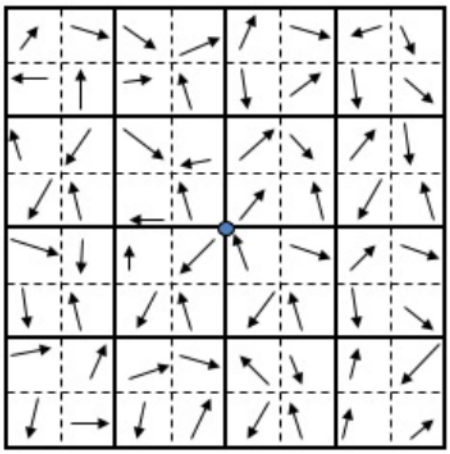


Figure 10. Un point clé (en bleu au centre de la grille) et les gradients d'images associés à sa région de (16 × 16) pixels.

Pour chaque sous-zone, les orientations de gradient θ sont réparties dans un histogramme d'orientations, divisant la plage des orientations en 8 intervalles uniformes correspondant à un voisinage V8 (**Figure 11**). Dans cette distribution, l'orientation θ du gradient définit la plage à incrémenter pondérée par son amplitude m.

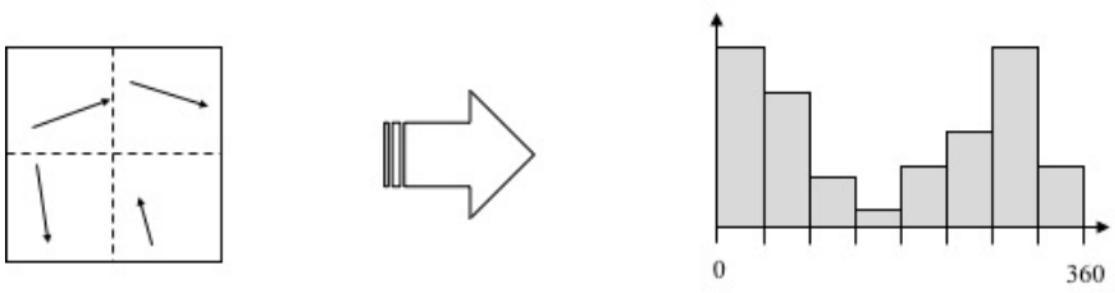


Figure 11. Exemple d'un histogramme d'orientation pour une sous-zone de 4 pixels par 4 pixels

Ainsi, nous obtenons 16 histogrammes des orientations avec 8 intervalles chacun. Par conséquent, localement une zone définissant un point clé est décrite par un vecteur de dimension 128.

Comme décrit dans [Lowe04], le descripteur SIFT présente des propriétés intéressantes et en particulier pour la description d'images représentant des zones urbaines [SiUn09]. En effet, il a été démontré [Lowe04] que ce descripteur local est invariant aux changements d'échelle, aux variations d'orientation, de point de vue et partiellement aux changements d'éclairage. De ce fait, le descripteur SIFT semble répondre de manière optimale aux contraintes de notre contexte d'application de reconnaissance de bâtiments dans de scènes urbaines.

II.1.2.2. Descripteur SURF

Le descripteur SURF (*Speed Up Robust Features*) [BoMG08], [PaPS13], [BaTG06] se présente comme une alternative de haute performance au descripteur SIFT. L'objectif est de diminuer la taille des descripteurs afin de réduire les temps de calcul afférents (aussi bien au niveau de l'extraction du descripteur qu'à celui de calcul des mesures de similarité) tout en améliorant l'assignation des orientations.

Pour déterminer les orientations dominantes on exploite une technique de filtrage par ondelette de Haar [ViJo01] dans un voisinage circulaire autour du point d'intérêt qui permet de définir des fenêtres de détection simples (**Figure 12**).

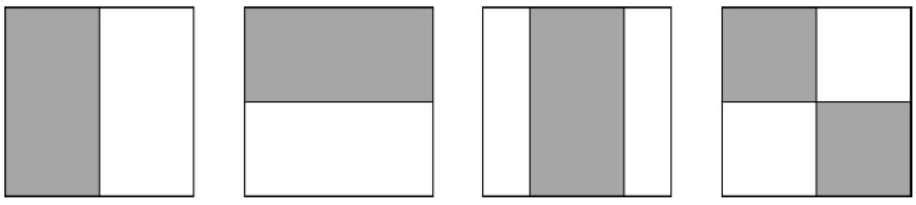


Figure 12. Exemple de fenêtres de détection définies par la méthode de Haar.

Les intensités des pixels dans un rectangle ainsi défini sont sommées à l'aide des images intégrales. Une image intégrale a la même taille que l'image originale. Sa valeur dans un point donné

(x, y) est par définition la somme de l'ensemble des valeurs de l'image originale précédent le pixel courant par rapport à un balayage de type raster scan, comme décrit dans l'équation (8) :

$$I_\Sigma(x,y) = \sum_{i=0}^{i<x} \sum_{j=0}^{j<y} I(x,y) \tag{8}$$

L'utilisation des images intégrales permet de réduire le nombre de calculs pour obtenir la somme des valeurs dans un rectangle. La différence de somme des intensités entre chaque région constitue une caractéristique de l'image définie par [ViJo01].

Ces masques sont appliqués pour chaque position de l'image, en commençant par des régions de petites tailles et allant de plus en plus grandes. La réponse des ondelettes de Haar est représentée selon deux vecteurs, l'un étant la réponse horizontale, l'autre la réponse verticale. L'orientation dominante est assignée à un point d'intérêt suivant la somme des réponses de l'ondelette de Haar dans la fenêtre créée précédemment.

Dans un second temps, le descripteur SURF est extrait à partir d'une région de 20×20 pixels subdivisée en 4×4 sous-régions au voisinage du point d'intérêt et orientée selon l'étape précédente. Le descripteur SURF est à l'instar du descripteur SIFT une méthode de description locale des points d'intérêts. Les descripteurs SURF sont plus rapides que les descripteurs SIFT. En revanche, ils sont moins riches en information et deviennent plus sensibles aux changements d'échelle, de rotation et d'illumination [PaPS13], [KhMW11]. Notre travail ayant pour cadre un environnement public et extérieur, la sensibilité aux changements d'illumination est un point faible non négligeable de ce descripteur. Pour cette raison, nous ne l'avons pas retenu dans nos travaux.

II.1.2.3. Descripteur HOG

La méthode de description par Histogramme de Gradients Orientés (HOG – *Histogram of Oriented Gradients*) a été introduite dans [VKMT13] et [VeFe15]. Il s'agit de calculer des histogrammes locaux de l'orientation du gradient sur une grille dense. La méthode a été initialement conçue pour la détection de silhouettes d'individus.

Localement, un objet dans une image peut être décrit par la distribution de l'intensité du gradient ou la direction des contours. Pour cela, une grille dense est appliquée à toute l'image, ce qui correspond à un découpage en plusieurs régions rectangulaires. Il s'agit alors de calculer pour chaque région l'histogramme des directions du gradient pour l'ensemble des pixels de cette région. L'ensemble de ces histogrammes forme le descripteur d'image HOG, normalisé par la suite. Pour un vecteur v de k-norme $\|v\|_k$, le facteur de normalisation f est défini selon la norme choisie comme décrit par les équations (9), (10) et (11) où ϵ est réel positif suffisamment petit :

Pour la norme L2 :

$$f = \frac{v}{\sqrt{\|v\|_2^2 + \epsilon^2}} \tag{9}$$

Pour la norme L1 :

$$f = \frac{v}{\|v\|_1 + \epsilon} \tag{10}$$

Pour la norme L1-racine :

$$f = \sqrt{\frac{v}{\|v\|_1 + \epsilon}} \tag{11}$$

Le descripteur HOG fournit une représentation globale, adaptée à un objet d'intérêt particulier. Néanmoins, le rapport d'aspect de l'objet est un paramètre qui peut influer directement sur les performances. Cela peut représenter un handicap dans notre cas, puisque les bâtiments à reconnaître peuvent présenter de tailles et de formes diverses.

II.1.2.4. Descripteur GIST

Le descripteur GIST [OlTo01] s'appuie sur la notion de *géon* [Bied87], qui représente une forme volumétrique simple, utilisée comme élément atomique de description d'un objet. Une représentation holistique d'une scène donnée peut être obtenue en cumulant l'ensemble des *géons* qui y peuvent être associés. Cette approche définit l'idée de « forme d'une scène » et la considère comme une entité entière, globale, alors qu'habituellement, une scène est vue comme un ensemble d'objets et son interprétation consiste en la détection et la localisation de ces objets.

Bien que l'idée de distinguer les objets dans une scène permet d'identifier un bâtiment recherché dans une image, il ne s'agit pas de distinguer chaque objet dans l'image mais bien de rechercher les images représentant le même bâtiment d'intérêt.

Le descripteur GIST met en œuvre le produit d'un filtre de Gabor g défini dans l'équation (12) et d'une enveloppe gaussienne. Dans l'équation (12), ρ et ϕ sont deux orientations définissant l'harmonique complexe pour un point x. Dans la partie définissant l'enveloppe gaussienne, μ est l'espérance mathématique et σ désigne l'écart type. Finalement, le descripteur GIST d'une image donnée est la convolution de cette image par différents filtres de Gabor dans 4 espaces et 8 directions aboutissant ainsi à 32 cartes de même taille que l'image originale regroupant différents points d'intérêts locaux. Chaque carte est par la suite partitionnée grâce à une grille de 4×4 donnant ainsi 16 régions par carte dans lesquels la moyenne des valeurs des points d'intérêt est calculée. La concaténation de ces 16 valeurs moyennes pour les 32 cartes résulte en le descripteur GIST de dimension 512. Ce descripteur global représente l'information apportée par le gradient de l'image (échelles et orientations).

$$g(x) = \exp(2j\pi\rho x + \phi) \cdot \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right) \tag{12}$$

Quels que soient les descripteurs utilisés pour la représentation d'images/scènes/objets, une étape de classification est nécessaire afin de pouvoir associer des labels sémantiques aux contenus visuels analysés. Détaillons-en à présent le principe.

II.1.3. Classification d'information

Une fois l'information extraite des images et décrite par des descripteurs, il s'agit de construire un vocabulaire et de répartir les images similaires dans des différentes classes/catégories disponibles.

Deux grandes familles d'algorithmes d'apprentissages peuvent être considérées, suivant si les classes sont prédéterminées (classification supervisée) ou non (classification non supervisée).

II.1.3.1. Classification supervisée

Si les classes sont définies au préalable, la classification consiste à distribuer les données dans ces différentes classes, selon des critères de proximité (distances ou mesures de similarité) dans l'espace des descripteurs.

Une première étape dite d'entraînement prend en considération une partie des données connues, réparties dans les classes prédéfinies de façon à construire une marge de séparation entre les différentes classes. Dans les cas les plus simples, la marge de séparation peut être considérée comme un hyperplan dans l'espace multidimensionnel des descripteurs considérés. Néanmoins, la complexité des données auxquelles nous avons affaire en pratique ne permettent pas dans tous les cas d'obtenir une séparation linéaire. Les techniques avancées [Suyk01], [YYGH09] conduisent à des marges de séparation non-linéaire, qui permettent d'isoler les différentes classes de manière efficace.

La deuxième étape, dite de test ou de prédiction, permet de déterminer la classe associée à une donnée inconnue, comme illustré **Figure 13**.

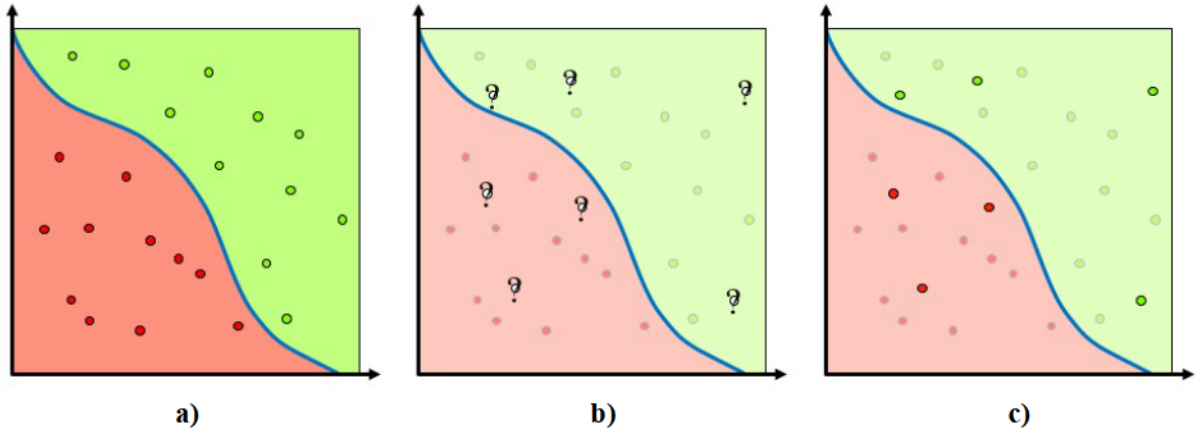


Figure 13. Exemple de classification supervisée **a)** Phase d'entraînement des données connues en deux classes (vert et rouge) avec construction de la marge de séparation (en bleu) ; **b)** Phase de test lors de laquelle les points inconnus (en blanc) sont attribués à l'une des deux classes ; **c)** Résultats des prédictions de classification.

Différents types de méthodes de classification supervisée sont aujourd'hui disponibles. Parmi les plus populaires, citons les méthodes de régression, le *boosting* [Scha90], les machines à vecteurs de support (*Support Vector Machines* – SVM) [BoGV92], les réseaux de neurones [SiVZ13], la méthode des *k* plus proches voisins (*k-nearest neighbors, kNN*) [Altm92] ou encore les arbres de décision [Quin87].

II.1.3.2. Classification non-supervisée ou *clustering*

Dans le cas où les classes dans lesquelles regrouper les données ne sont pas définies *a priori*, la classification est non-supervisée. Avant de distribuer les images dans différentes classes, les données sont partitionnées dans le but de créer des grappes ou clusters, comme illustré **Figure 14**.

La méthode des *k*-moyennes représente une des approches de *clustering* les plus populaires [ChFr15], [CAPZ15]. Il s'agit d'une méthode de partitionnement en apprentissage non supervisé, le but étant de regrouper les données en *k* ensembles en minimisant la distance entre chaque point d'une même partition. La procédure commence par distribuer *k* points représentant la position moyenne de chaque ensemble initial. Ces points peuvent être choisis initialement au hasard. Par la suite, on assigne chaque donnée à la partition la plus proche. Une fois toutes les données distribuées, la moyenne de chaque cluster est mise à jour et les données redistribuées suivant les nouvelles moyennes et nouveaux ensembles définis jusqu'à ce qu'il n'y ait plus de changement de classe (convergence).

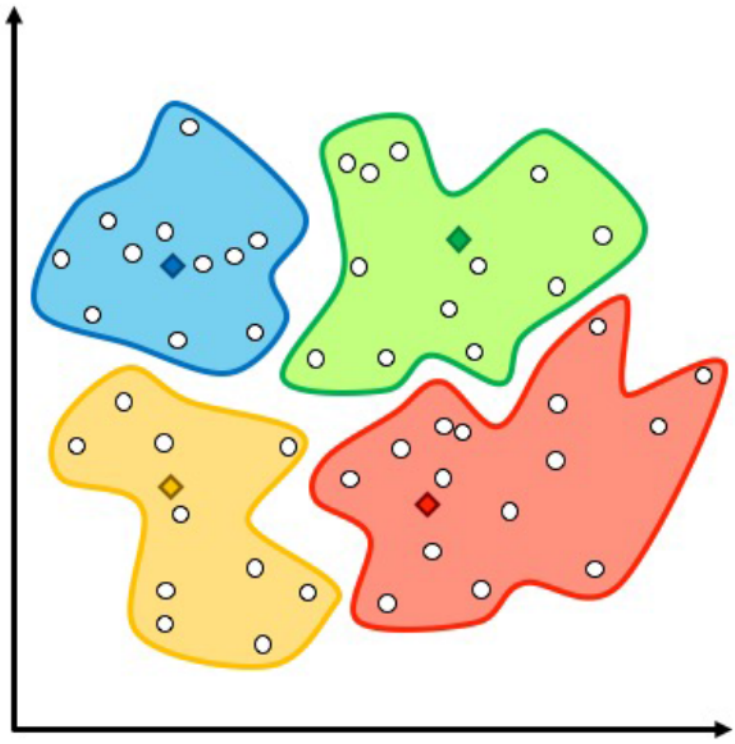


Figure 14. Exemple de classification non-supervisée. Les points à répartir sont en blanc, les losanges de couleurs représentent les centroïdes des *clusters* de la couleur correspondante.

La classification supervisée permet un apprentissage automatique pour la reconnaissance d'objets. Cette approche permet ainsi de regrouper les images dans la base de données suivant les objets qu'elle contient.

D'autre part, ce modèle de classification permet la construction de vocabulaire par partitionnement des descripteurs pour faciliter leur indexation. Cela correspond à la dernière étape de construction du vocabulaire utilisé comme illustré **Figure 15** par la suite lors de la phase de recherche réelle d'images illustrée **Figure 16**.

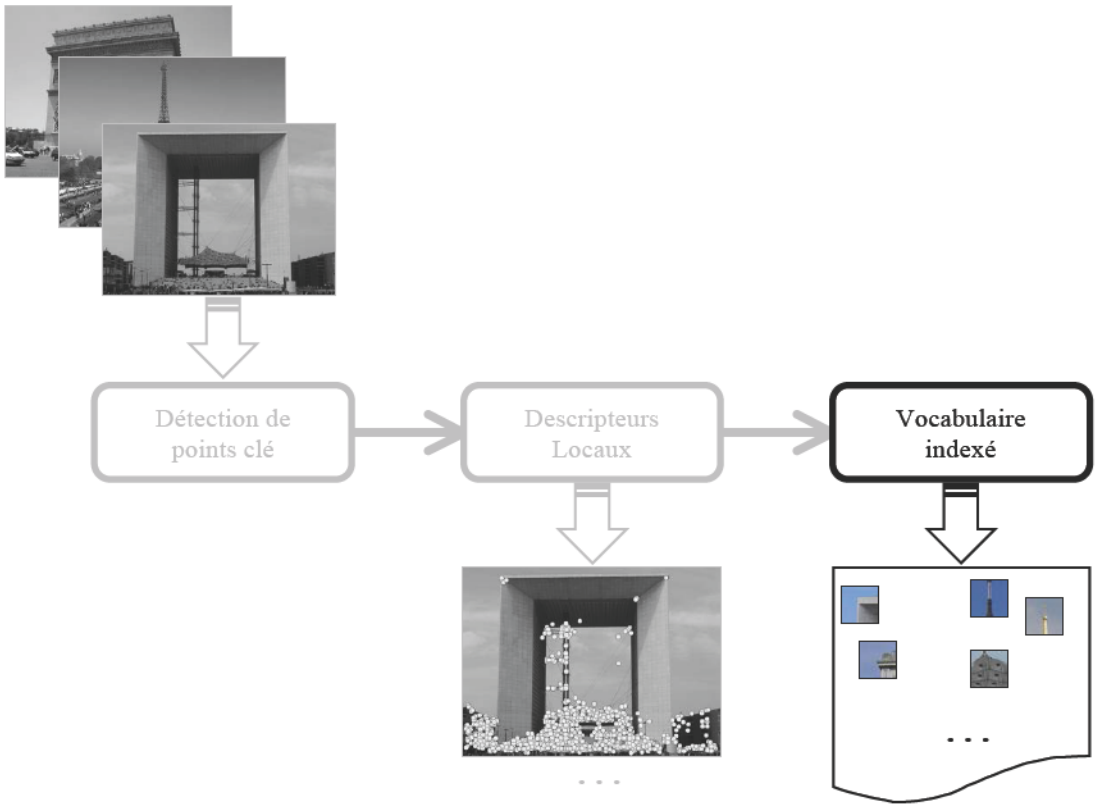


Figure 15. Étapes de classification lors du processus de construction de vocabulaire indexé.

Les deux prochains paragraphes présentent la dernière étape de notre processus de recherche d'image illustrée à travers les modèles de BOW [SiZi03] et VLAD [JDSP10] permettant de retourner les images similaires à une image requête par ordre de similarité.

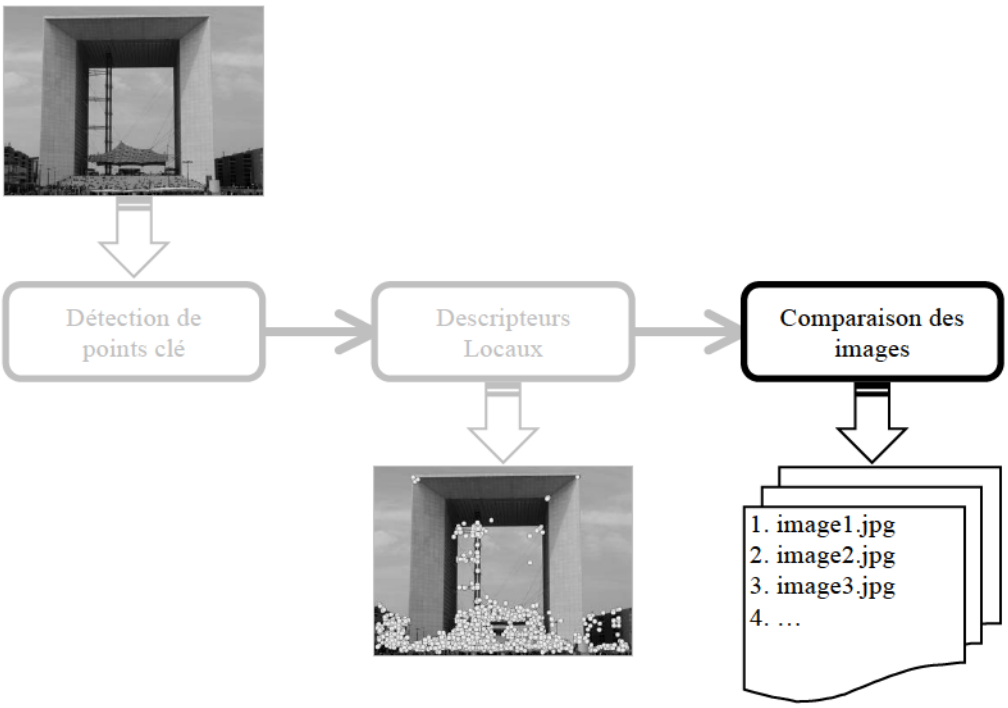


Figure 16. Dernière étape de reconnaissance d'images.

II.1.4. Modèle *Bag of Words*

Une des méthodes les plus populaires utilisée pour la recherche d'images est le modèle *Bag of Words* (BOW) [SiZi03], qui s'inspire des techniques de recherche textuelle tout en les étendant au domaine image. Ainsi, les mots textuels sont ici remplacés par des « mots visuels », qui correspondent aux descripteurs associés aux points clés détectés, quantifiés selon un dictionnaire de prototypes.

L'objectif est donc de créer un index de mots visuels pour chaque image à analyser. Le processus de construction du dictionnaire des prototypes est une étape clef pour la mise en œuvre de ce processus. Par exemple, dans le cas des descripteurs SIFT, des milliers de descripteurs de points clés de dimension 128 sont extraits pour chaque image. Pour des bases de données de plusieurs milliers d'images, cela conduirait à un vocabulaire de millions de mots, chacun exprimé dans un espace de dimension 128. A l'évidence, il n'est pas possible de comparer chacun des points d'une image aux millions d'autres descripteurs. Le principe consiste consiste à utiliser un vocabulaire plus petit, qui est construit en utilisant une méthode de clustering inspirée de la méthode des *K*-moyennes décrite dans [PCIS07]. Cela revient à effectuer une quantification vectorielle de l'espace des descripteurs pour regrouper les descripteurs similaires (au sens de la métrique caractéristique de l'espace des descripteurs considérés) dans des *clusters*. Chaque *cluster* sera ensuite représenté par un seul prototype (habituellement le centroïde du cluster considéré). De cette façon, il est possible de définir une taille spécifique de vocabulaire indépendamment de la méthode utilisée pour extraire des caractéristiques locales. Tous les mots visuels les plus proches sont regroupés en *K clusters* avec la méthode de classification des *K*-moyennes où *K* correspond au nombre de clusters désirés et donc à la taille voulue pour notre vocabulaire.

A partir de ce vocabulaire de prototypes visuels, le modèle BOW repose sur la construction d'un histogramme qui permet de comptabiliser la fréquence relative d'apparition de tous les mots

visuels du vocabulaire dans chaque image (**Figure 17**). Par la suite, une image peut être complètement représentée par ses histogrammes de fréquences associés pour chaque point clé extrait de cette image.

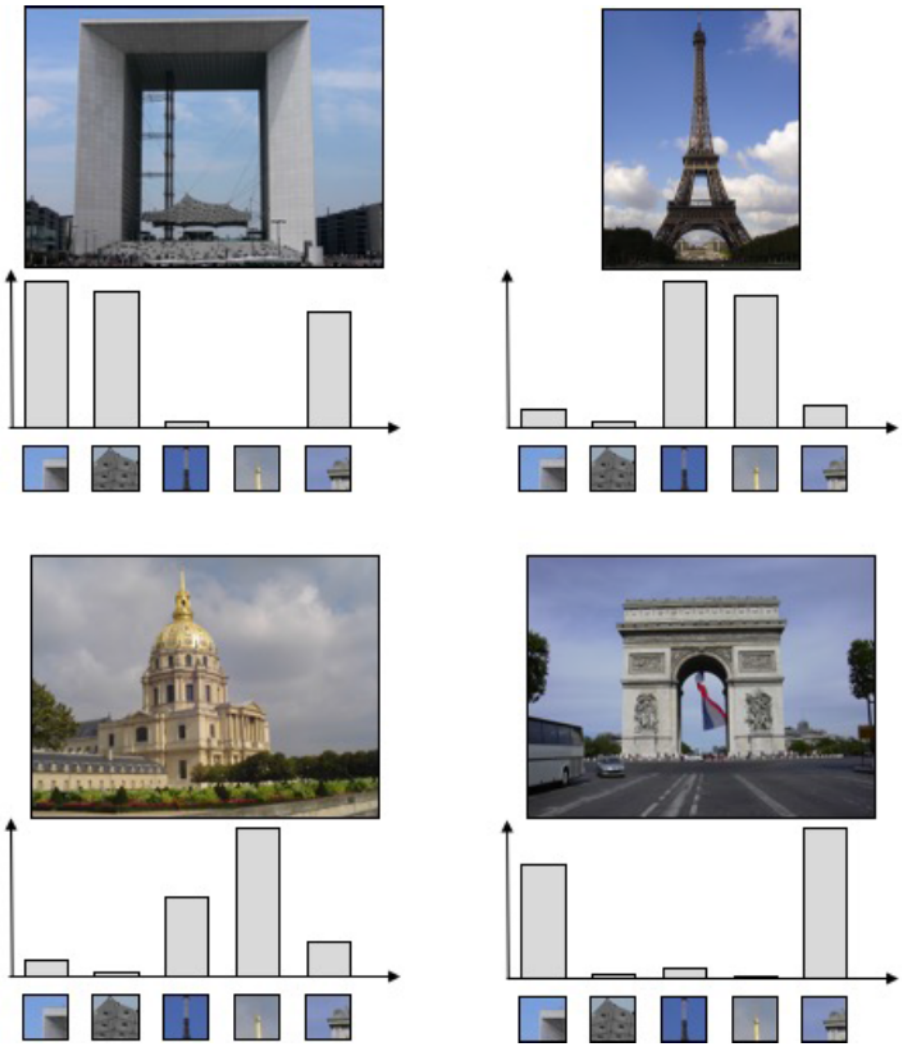


Figure 17. Exemple d'histogramme des fréquences dans la base de données Paris6k pour quatre images différentes de quatre classes différentes et un vocabulaire de cinq mots.
a) La Défense **b)** La Tour Eiffel **c)** Le Sacré Cœur **d)** L'Arc de Triomphe. Les deux premiers mots visuels dans les histogrammes font partie de l'image de La Défense, puis un mot visuel est extrait de chacune des trois images restantes pour cet exemple. Nous pouvons remarquer dans cet exemple que le coin à la fois dans l'arche de La Défense et de l'Arc de Triomphe sont similaires visuellement ce qui est reporté par des fréquences élevées pour ces deux mots dans les deux images correspondantes.

En s'inspirant également des méthodes d'indexation textuelle, le modèle BOW adopte une pondération de type *tf-idf* (*term frequency – inverse document frequency*), qui permet de gérer à la fois l'influence des mots trop communs (*i.e.*, qui apparaissent dans une majorité de documents) qui sont, par définition, peu discriminants et celle des mots spécifiques à certains documents, qui permettent leur identification.

Le terme *tf* (*term frequency)* représente le nombre de fois où le mot apparaît dans le document par rapport au nombre total de mots dans le document, comme décrit dans l'équation (13).

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{13}$$

où i représente le mot de requête, j le document étudié et n_{ij} est le nombre d'occurrences du terme i dans le document j.

Les mots trop communs ne sont pas discriminants pour retrouver des documents et donc ne contribuent pas à les identifier les uns des autres. A l'opposé, les mots spécifiques apparaissant peu souvent donnent plus de sens et plus de puissance discriminative au document donné. La prise en compte de ces aspects est réalisé avec l'aide du facteur de fréquence inverse de document (en anglais *inverse document frequency* ou *idf*), défini dans l'équation (14) :

$$\mathrm{idf}_i = \log \frac{|D|}{\left|\{d_j : t_i \in d_j\}\right|} \tag{14}$$

où $|D|$ est le nombre total de documents dans le corpus et $\left|\{d_j : t_i \in d_j\}\right|$ le nombre de documents dans lesquels le mot t_i apparaît.

Pour des mots communs, apparaissant dans un grand nombre de documents le terme *idf* prend des valeurs relativement faibles. Au contraire, pour des mots spécifiques, caractéristiques à des sous-ensembles de documents les valeurs du facteur *idf* seront plus importantes.

Enfin, le produit de ces deux statistiques correspond à la discriminance d'un mot par rapport à un document dans une collection de documents. Il s'agit du score *tf-idf* (*term frequency-inverse document frequency*) (15), utilisé pour pondérer tous les termes du vocabulaire et construire ainsi un index plus efficace.

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \cdot \mathrm{idf}_i \tag{15}$$

Par la suite, lorsqu'il s'agit de rechercher les images correspondant à une image requête, il suffit de comparer les histogrammes de fréquences de l'image.

La mesure de similarité entre deux "mots" visuels peut être calculée par une distance euclidienne (16) ou cosinus (17) classiques entre les histogrammes calculés dans les images de la base

de données en utilisant le score de *tf-idf* pour chaque mot. Ce processus conduit à une liste d'images classées par similitude décroissante par rapport à l'image requête.

$$\Delta(u,v) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2} \tag{16}$$

$$\cos(u,v) = \frac{u \cdot v}{\|u\| \cdot \|v\|} \tag{17}$$

II.1.5. Descripteurs de Fisher

Le descripteur de Fisher [PLSP10] est une représentation d'image obtenue par un processus d'agrégation d'informations locales. Le noyau de Fisher [JaHa99] représente une mesure de similarité entre deux vecteurs, dits vecteurs de Fisher [PeDa07], [PeSM10]. Ces vecteurs sont obtenus à l'aide d'une analyse statistique de l'image par maximum de vraisemblance par rapport à un modèle probabiliste donné et paramétré selon un vecteur de paramètres λ.

Ainsi, lors d'une première étape d'entraînement, en utilisant le principe de maximum de vraisemblance, le paramètre λ du modèle est défini de façon à correspondre aux données d'entraînement dans l'équation de probabilité conditionnelle $p(I|\lambda)$ de l'image I. Finalement, une image est entièrement décrite par un vecteur de dimension $2N$, c'est-à-dire deux fois la taille du vocabulaire. Cependant, malgré la diminution de la taille des données, [PeSM10] montre que moins d'une image sur deux est correctement retrouvée avec ce modèle de recherche d'image.

II.1.6. Vecteurs VLAD

Dans [JDSP10], les auteurs proposent une autre représentation d'images, appelée VLAD (*Vector of Locally Aggregated Descriptors*). Il est présenté dans [ArZi13] comme une amélioration du modèle de représentation BOW qui permet d'obtenir une représentation plus compacte.

A l'instar du descripteur de Fisher présenté précédemment, un vecteur VLAD est la représentation globale des différents points d'intérêt locaux d'une image. Comme pour le modèle BOW [SiZi03], la représentation VLAD [JDSP10] est fondée sur un vocabulaire de caractéristiques locales. Cependant, au lieu de regrouper tous les descripteurs en histogrammes de fréquences, on ne considère que les *vecteurs résiduels*. Autrement dit, au moment de construire le vocabulaire en regroupement les descripteurs locaux extraits, au lieu de compter le nombre de descripteurs dans chaque cluster comme dans le modèle de BOW, il s'agit de quantifier et prendre en compte la différence entre les descripteurs attribués à un cluster et le centre de ce cluster.

Tout d'abord, à partir d'un ensemble de N descripteurs locaux SIFT $X = \{x_1, \dots, x_n\}$ extraits d'une image, un vocabulaire $C = \{c_1, \dots, c_k\}$ est construit en regroupant les n descripteurs, en k clusters. L'objectif est d'attribuer un descripteur local x à son cluster le plus proche de centre c, comme décrit dans l'équation (18).

$$x \mapsto n(x) = \arg\min_{c \in C} \|x - c\|^2 \tag{18}$$

Avec $n(x)$ le centre du cluster le plus proche attribué au descripteur local x.

Avec ce vocabulaire, pour chaque mot visuel c_i, les différences $x - c_i$ sont sommées dans un sous vecteur v_i comme décrit dans l'équation (19). Comme ce sous-vecteur v_i est calculé avec la somme des différents descripteurs locaux, sa dimension est identique à celle du descripteur local. Ici, comme nous utilisons le descripteur local SIFT, la dimension de chaque sous-vecteur v_i est donc de 128.

$$v_i = \sum_{j \in [1,n]} x_j - c_i \tag{19}$$

où x_j est le descripteur local affecté au cluster de centre c_i .

La concaténation de tous ces sous-vecteurs v_i pour l'ensemble des mots visuels du vocabulaire considéré représente le vecteur VLAD $v = [v_i \dots v_k]$. Étant donné que les sous vecteurs v_i sont tous de dimension n (où n est la dimension des descripteurs locaux), et qu'il y a k mots du vocabulaire, la dimension du vecteur VLAD est de $n \times k$.

Afin de compacter encore plus la représentation, dans [DGJP13] les auteurs proposent d'appliquer une réduction de dimension des descripteurs, à l'aide d'une méthode d'analyse en composantes principales (ACP). L'ACP est dans ce cas appliquée avant l'agrégation des vecteurs résiduels, dans l'espace des descripteurs SIFT. Seuls les coefficients correspondant aux 64 premières directions principales sont ainsi retenus.

La méthode VLAD conduit à une représentation d'image qui s'appuie sur le même vocabulaire que le modèle BOW mais avec une dimension beaucoup plus faible.

Dans nos expérimentations, nous avons retenu pour évaluation et comparaison les deux modèles BOW et VLAD.

Analysons à présent comment ces différents modèles de représentation sont appliqués à des objectifs de reconnaissance de bâtiments.

II.2. Reconnaissance d'images de bâtiments

II.2.1. Recherche d'images par descripteurs SIFT et modèle BOW

Les descripteurs SIFT sont utilisés dans un contexte de *street-view* [PISZ10] dans le cadre du projet *iTowns* à Paris, dont les objectifs sont de retrouver les différentes vues d'une scène urbaine à partir d'une image requête.

La méthode s'appuie sur des comparaisons point par point entre l'image requête et chaque image de la base de donnée. Par conséquent, la procédure de comparaison est linéaire et séquentielle, image après image. Les performances afférentes dépendent fortement de la taille de la base de données.

Il est alors proposé une approche par vote. Chaque point-clef de l'image requête vote ou non pour une image dans la base de données, suivant si les distances entre le point-clef de la requête et ses deux plus proches voisins sont inférieures à un seuil donné. Le nombre de votes par image définit ainsi le rang de l'image dans la reconnaissance. Cependant, les points clefs de l'image risquent d'être reliés à de faux résultats car ils sont indépendants du contexte. Ainsi, une image hors sujet risque d'obtenir un haut rang de façon aléatoire. Pour pallier à ce problème il est proposé par [PISZ10] une méthode de vérification de la cohérence géométrique à base de la méthode RANSAC (*RANdom SAmple Consensus*) [FiBo81].

La méthode de BOW est utilisée et améliorée par [BoMG08] dans le contexte de la localisation d'un robot par la reconnaissance de la scène qui l'entoure. Ici, les descripteurs SURF sont utilisés comme éléments locaux. Les expérimentations ont été conduites sur une base de données [NiSt06] de 6 376 images où chaque objet est représenté quatre fois, avec des poses différentes. Il est aussi proposé de construire un vocabulaire hiérarchisé. Au lieu de comparer linéairement tous les mots visuels, le vocabulaire est partitionné autour d'un petit nombre de mots et chaque cluster est repartitionné pour former des sous-vocabulaires.

II.2.2. Reconnaissance d'images pour la géolocalisation

Dans [GOSP13], les auteurs proposent une technique visant à reconnaître des images pour localiser géographiquement une scène grâce à des marqueurs géographiques (*géotags*). Cette méthode est inspirée de la reconnaissance utilisant la méthode de classification SVM. Etant donné que peu de résultats positifs sont obtenus lors de cette phase d'apprentissage, une nouvelle approche d'apprentissage utilisant les résultats négatifs est alors proposée et expérimentée sur une base de données de 25 000 images de rues de Pittsburgh. La méthode adopte la représentation de BOW et les descripteurs utilisés sont des SURF ou des SIFT, normalisés en norme L_2. Les similarités entre images sont calculées avec un produit scalaire, ce qui permet une certaine robustesse au bruit dans l'arrière-plan et aux occlusions partielles.

D'autres travaux [PISZ10] et [MPCM10] cherchent à améliorer la précision et la vitesse de calcul du modèle de BOW en utilisant une structure géo-marquée d'images localisées sur une carte et décrivant le monde réel 3D. L'idée de graphe d'images est décrite dans [PhSZ11], [TuLo09] et

[ToSP11]. Dans ce graphe, chaque nœud est une image, reliée à des images identifiées comme similaires.

D'autres approches [KnSP10] et [ScBS07] utilisent des données géo-marquées pour l'apprentissage pour sélectionner les points-clefs locaux suivant la géolocalisation de l'image ou encore pour reclasser les résultats d'images après leur reconnaissance [ZaSh10]. Ainsi, il est possible de déterminer la localisation de l'image requête en récupérant la marque géographique de l'image la plus proche reconnue dans la base de données.

Les expériences réalisées par [KnSP10] arrivent à atteindre un pourcentage de requêtes correctement localisées pour différents lieux de 48%, en quasi temps-réel avec une vérification spatiale sur 17 000 images de *Google Street View* géo-marquées. Cela montre bien que la reconnaissance de lieux dans le domaine de l'imagerie peut encore être améliorée.

Les résultats présentés dans [ScBS07] ont été obtenus sur une base de données de 10 millions d'images prises à partir d'un véhicule dans une grande ville et marquées de leurs coordonnées GPS. Comme résultat marquant, notons que les 10 premiers résultats retrouvés après requête sont correctement localisés à 80% en vérifiant la consistance géométrique des points d'intérêts et dans des temps pouvant descendre jusqu'à 0,2 secondes par requête. Dans ces conditions, un résultat est considéré comme correctement localisé s'il se trouve à moins de 10 m de la vérité de terrain.

Les tests conduits par [ZhTM13] se basent sur 521 images pour former une petite base de données. Les résultats montrent que 60% de l'ensemble de test est correctement localisé (à moins de 100 m de la vérité de terrain) mais que cette méthode souffre beaucoup d'occlusion du bâtiment par des objets non-permanents (voitures, personnes, …) ou avec peu d'information (arbres, buissons, …). Cependant, [ZaSh10] propose une recherche par groupe d'images (un groupe étant composé de 3 images dans ce cas, distantes de moins de 300 m dans un même groupe). Dans ce cas, la localisation correcte est obtenue dans plus de 90% des cas et il est montré dans [ZaSh10] que plus le groupe d'images requête est important, plus le résultat sera précis, c'est-à-dire plus proches de la localisation géographique réelle.

Pour une application de navigation basée sur la « vision » d'un robot, les auteurs de [GoTG04] présentent différentes contraintes apportées par la reconnaissance de bâtiment en extérieur dans un environnement public. Les points de vue et condition d'éclairage peuvent dans ce cas varier drastiquement et la végétation environnante représente un obstacle supplémentaire. Il s'agit alors de mettre en œuvre des descripteurs d'information locaux invariants aux transformations affines.

L'importance d'une reconnaissance de bâtiments fiable et efficace pour une utilisation de géolocalisation est également montrée par [ZhKo07]. Dans ce cadre, les auteurs proposent une approche de reconnaissance hiérarchisée en plusieurs étapes. Dans un premier temps, les images sont indexées selon la localisation d'histogrammes de couleur. Cette approche s'appuie sur le fait que les bâtiments sont visuellement reconnaissables par leurs lignes parallèles et orthogonales dans un repère présentant des lignes de fuites. Seuls les pixels dont l'orientation du gradient est dans la même direction que ces lignes de fuites sont donc pris en considération pour interpréter la distribution colorimétrique de l'image. La deuxième étape de reconnaissance de bâtiment évalue les descripteurs SIFT des images retenues dans la première étape pour affiner la reconnaissance. Cependant, cette méthode a pour prérequis le fait que chaque image ne représente qu'un seul bâtiment pour que la première étape de reconnaissance soit pertinente lors de l'indexation par localisation des clusters de couleurs.

II.2.3. Utilisation d'informations géo-localisées

D'autre part, la géolocalisation peut être également exploitée pour améliorer les performances des algorithmes de recherche d'image. Ainsi, les tags géographiques des images de la base de donnée sont exploités comme dans les exemples de [BiMS09], [ArZi13], [KnSP10], [ScBS07], [ZaSh10], [JGZY11] et [ZZSA09]. Cependant, ces métadonnées pourraient aussi être utilisées pour créer des classes de localisation dans la base de données. Lors d'une requête avec un *smartphone*, le GPS peut alors être mis à contribution pour identifier la position actuelle de l'utilisateur qui servira par la suite à filtrer la base de données d'image *a priori*, contrairement aux utilisations actuelles du GPS qui servent de vérification après avoir obtenus les résultats.

Une première étape consiste donc à comparer cette position avec les clusters géographiques de la base de données et ainsi obtenir un classement des clusters les plus proches géographiquement de la position de l'utilisateur. Une fois le ou les meilleurs regroupement(s) ainsi identifiés, il est alors possible de créer une nouvelle base de données (*i.e.*, une sous-base de données de l'originale), en ne gardant que les monuments et bâtiments proches de la position géographique de l'utilisateur. La position géographique apportée par le GPS du *smartphone* de l'utilisateur est donc une nouvelle donnée et information discriminante permettant de réduire la taille de la base de données d'image.

Comme rapporté par [KnSP10], [ScBS07] et [ZaSh10] la méthode de localisation GPS n'est pas toujours très précise. Cependant, une localisation approximative permettrait néanmoins de pouvoir sélectionner la ou les classes de la base de données d'image qui sont les plus proches géographiquement de la zone géographique où se situe l'utilisateur. Ayant ainsi sélectionné un nombre réduit d'images et obtenu une plus petite base de données, il est possible d'utiliser des méthodes plus lourdes en temps de calculs. A l'instar de [KnSP10] qui utilise le GPS après la reconnaissance d'images pour reclasser les résultats obtenus afin d'obtenir une meilleure cohérence géographique, il s'agit ici d'utiliser cette position *a priori*, dans le but de diminuer la part de calcul et de temps dédié à la reconnaissance d'images. Il serait alors possible de réaliser un apprentissage spécialisé pour les quelques images filtrées. La méthode de BOW semble toujours adaptée à cette reconnaissance en rejetant avant l'étape de reconnaissance les descripteurs n'appartenant pas aux images de la classe géographique isolée.

La méthode introduite par [ZZSA09] propose de recenser différents monuments historiques à travers le monde à partir d'images agrémentées d'informations GPS et touristiques issues d'Internet. Cette base de données ainsi construite est complétée en recherchant les images similaires aux premières déjà répertoriées. Cette construction nécessite néanmoins une validation après avoir regrouper les différentes images en *clusters* selon leur localisation géographique et les différentes informations touristiques obtenues.

II.2.4. Système d'Information Géographique

Un Système d'Information Géographique ou SIG est une base de données recueillant des données spatiales et géographiques. Ce système permet d'apporter une information sémantique à l'information visuelle, comme proposé par [DuZZ15], dans le but d'affiner la classification des différentes catégories d'une base de données de façon sémantique. L'approche proposée se montre

efficace pour classifier des bâtiments reconnus dans des images satellites. Dans ce cas, les bâtiments peuvent en effet être visuellement similaires les uns aux autres.

Cependant, notre cadre est celui de photographies de monuments prises du point de vue d'un utilisateur lambda. Dans ce cas, bien que ce système puisse permettre d'affiner la classification des différentes catégories d'images, il s'agirait alors de construire une base de données complémentaire définissant les informations sémantiques des différents bâtiments. Ceci pourrait être fait en complément de notre reconnaissance de bâtiments mais ne semble pas apporter une information supplémentaire utile dans un premier temps.

II.2.5. Applications adaptées aux terminaux mobiles

La recherche d'image à partir de terminaux mobiles intelligents présente aujourd'hui un intérêt grandissant. En effet, les méthodes de recherches d'images dans des larges bases de données sont couteuses en mémoire et en calcul. La question qui se pose alors est comment les rendre efficaces dans un contexte de mobilité.

Dans [PISZ10], les auteurs montrent que lors de la recherche d'une requête dans une base de données, les descripteurs sont comparés un à un de façon linéaire et séquentielle. Ainsi, plus il y a de descripteurs, plus il y a d'information. La précision sera alors supérieure au prix d'un nombre plus important de comparaisons à effectuer et donc d'un temps d'exécution plus élevé. Aujourd'hui implémentées et testées sur des ordinateurs ayant les capacités suffisantes, ces méthodes doivent être adaptées pour pouvoir être utilisables sur un terminal mobile ayant des capacités de calcul/mémoire plus limités, tout en tenant compte des critères de temps de réponse et de précision pour l'utilisateur.

La reconnaissance d'objets permet à un terminal mobile de reconnaître son environnement et peut ainsi l'aider à se localiser comme expliqué par [BoMG08] et [AnTD05]. Les descripteurs SIFT et la représentation par BOW sont mis en œuvre dans le but de reconnaître les différents objets « vus » par le terminal mobile de façon entièrement automatique. Par reconnaissance des différents objets de la scène environnante, cela permet donc au terminal de se situer en utilisant les informations disponibles dans les images retrouvées dans la base de données.

Cependant, le problème de mémoire est soulevé par [GGHQ06] dans le cadre des terminaux mobiles. La robustesse au bruit, à l'occlusion, à la variation de luminosité et à la pose des objets est aussi souligné dans [GGHQ06]. Pour gérer ces aspects, les auteurs proposent l'utilisation de descripteurs SIFT.

En tout état de cause, l'objectif est de réduire la mémoire utilisée lors la reconnaissance d'un bâtiment dans une base de données.

Une méthode de combinaison d'information GPS et de reconnaissance visuelle d'images à partir de terminaux mobiles est proposée dans [HuMa05]. Pour chaque image de la base de données, l'information géo-localisée des bâtiments de la base de données est nécessaire afin de déterminer la position réelle d'acquisition de l'image requête que l'on cherche à déterminer. Cette application nécessite cependant au préalable une reconnaissance visuelle fiable du bâtiment utilisé en référence pour la géolocalisation de l'utilisateur.

Dans [LHSA14], il est proposé d'utiliser la reconnaissance de bâtiments par un terminal mobile pour faciliter la géolocalisation sans informations GPS. La première étape de l'approche est de filtrer les images de la base de données selon leur couleur. Dans un second temps, des descripteurs

locaux sont extraits des images résultantes et comparés à l'image requête. Enfin, une comparaison de la pose approximative de l'objet entre l'image requête et les images de la base de données permet de situer relativement l'utilisateur par rapport au bâtiment reconnu. Cependant, la méthode s'appuie sur l'hypothèse que chaque bâtiment de la base de données est répertorié selon différentes poses. D'autre part, la distance entre l'utilisateur et le bâtiment peut aussi induire une erreur de positionnement et une erreur de reconnaissance *a priori* si la reconnaissance des couleurs n'est pas suffisante. Enfin, pour cette méthode faisant intervenir des informations tant colorimétriques que locales, le temps de calcul sur un terminal mobile est relativement long, comme expliqué dans [KLCC06].

La solution proposée par [KLCC06] pour un système de navigation embarqué consiste à occulter les objets non pertinents dans la reconnaissance de bâtiments (*i.e.* la route, les véhicules, les arbres,...). La méthode utilisée ici exploite une reconnaissance de contours des objets d'une image. Appliquant un masque prédéterminé, il est alors possible de rejeter les voitures et la route. Les objets présentant des segments dont les directions sont trop aléatoires sont considérés comme étant des arbres et sont également rejetés. Ainsi, seules les zones sans véhicule, sans route et sans végétation sont conservées pour la reconnaissance de bâtiment.

Cependant, cette approche permet de reconnaître une classe d'objet (ici les bâtiments) mais sans en faire la distinction dans une même image. La méthode est donc applicable dans le cas où le monument recherché est isolé d'autres bâtiments, ce qui n'est pas toujours le cas dans un environnement urbain. L'idée principale soulevée par les auteurs est de rejeter les objets pouvant apporter une fausse information pour la reconnaissance d'un bâtiment et ainsi occulter les véhicules, végétations et autres interférences.

Une utilisation des descripteurs locaux SIFT dans une application de reconnaissance sur un terminal mobile est présentée par [FrSP06]. Les descripteurs locaux extraits sont regroupés en sous-espaces à l'aide d'une analyse en composantes principales. Dans cette étape de cartographie des descripteurs SIFT dans l'image donnée, un seuil prédéfini permet d'éliminer les descripteurs ambigus, *i.e.* non suffisamment discriminants ou trop redondants. En complément, un arbre de décision construit sur la base d'une estimation du maximum *a posteriori* permet de rejeter les descripteurs apportant le moins d'information. La méthode présentée montre bien l'avantage apporté par l'utilisation de descripteurs locaux apportant une information précise mais aussi la nécessité de sélectionner ces descripteurs locaux de façon à ne conserver que l'information nécessaire. Cependant, les expérimentations menées par [FrSP06] sont conduites dans un cadre spécifique dans lequel les images représentent uniquement des façades de bâtiments, le but étant alors de rejeter les descripteurs extraits de l'arrière-plan. La construction de l'arbre de décision est donc complexifiée lorsque plusieurs bâtiments sont représentés sur l'image.

II.2.6. Premier bilan

De cet état de l'art, il ressort que les modèles de BOW et son amélioration VLAD restent les méthodes les plus efficaces conciliant efficacité de temps de calcul et précision de la reconnaissance. Quant aux descripteurs, les choix les plus courants concernent les représentations SIFT ou SURF.

Etant donné qu'il est prévu de réduire la base de données et de la filtrer suivant la localisation géographique de l'utilisateur, le traitement de la reconnaissance pourrait s'effectuer sur un nombre réduit d'images. Ainsi, les descripteurs SURF pourraient être adaptés si la base de données est

suffisamment réduite. D'autre part, la sensibilité des descripteurs SURF aux changements d'illumination les pénalise dans un contexte où les images représentent des scènes extérieures.

Avec les approches présentées jusqu'à présent, toute l'information obtenue à partir de l'ensemble des images de la base de données est utilisée pour former le vocabulaire de mots visuels. Comme montré précédemment, pour les images acquises en extérieur et dans des espaces publics, des éléments indésirables présents dans les scènes (personnes, véhicules, végétation, bâtiments voisins, …) peuvent venir interférer avec l'objet d'intérêt. Ce problème est illustré **Figure 18** et est identifié également dans [LHSA14].

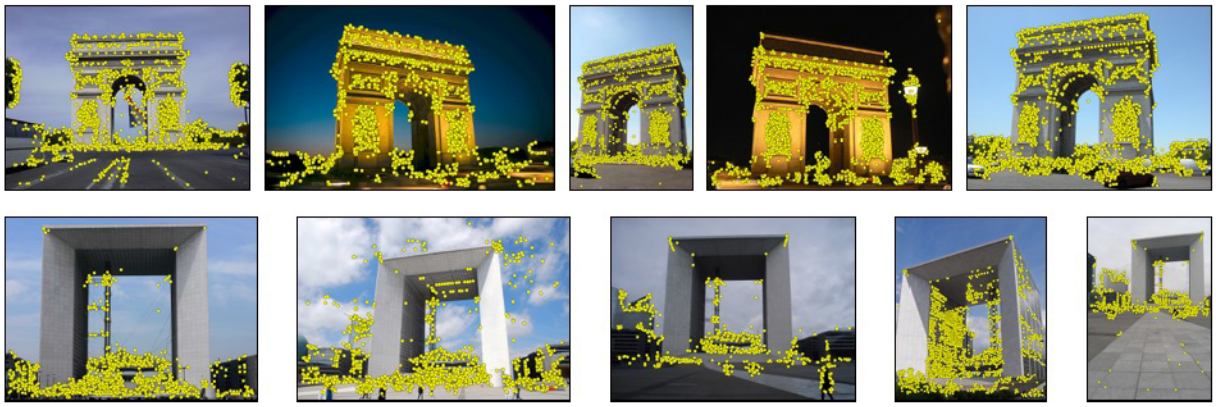


Figure 18. Exemples de points d'intérêts extraits sur des images représentant l'Arc de Triomphe et l'arche de la Défense dans la base de données Paris6k [Lowe04].

Afin de pallier à ces inconvénients, plusieurs solutions ont été proposées dans la littérature, qui visent à rendre le processus de recherche plus fiable. Les différentes familles d'approches sont décrites dans la section suivante.

II.3. Tendances et techniques émergentes

II.3.1. Utilisation d'information textuelle

Dans [LiCH09], [ChFr15] et [JLYL10]des images extraites d'Internet représentant des scènes urbaines sont considérées. Les auteurs soulignent la problématique de donner un sens à toutes ces données. L'objectif est ici de regrouper les images en fonction des bâtiments représentés en différentes classes. Cependant, le contexte de cette classification est non seulement basé sur le contenu visuel, mais fait aussi intervenir des labels et informations textuelles jointes à la photographie par les différents utilisateurs. Le problème principal est que les images sont étiquetées par des utilisateurs, avec des perceptions et sensibilités variés. Cette information textuelle ne peut donc pas être objective et encore moins uniforme de par la diversité des utilisateurs, ce qui la rend peu fiable.

II.3.2. Extension de requêtes

Un modèle mettant l'accent sur l'amélioration de l'information dans les images requêtes pour la recherche d'images est proposée dans [CMPM11], [PCIS08] et [ArZi12] et utilisé dans [PISZ10] et [ArZi13]. A partir d'un échantillon requête extrait d'une image représentant un bâtiment, le principe consiste à étendre automatiquement la région considérée pour la requête pour inclure entièrement l'objet. Cependant, si l'image requête est déjà considérée entièrement, la région ne peut pas être étendue. En outre, cette méthode d'extension de requête n'est traitée qu'après avoir obtenu les résultats de la recherche d'image et est utilisée dans le but d'améliorer le classement de ces résultats, dans le cadre d'un processus de *re-ranking*. Ainsi, si les images récupérées en premier lieu ne sont déjà pas détectées correctement, l'approche proposée par [CMPM11] relègue leur rang plus bas dans le classement des résultats, mais ce résultat aberrant n'est pourtant pas éliminé. Cependant, les auteurs montrent que la sélection des informations locales influence le sens concret donné aux objets dans un paysage urbain complexe. En trouvant la corrélation liant localement différents points d'intérêt, il est possible de choisir de façon plus précise et pertinente ceux décrivant un objet entier. Il s'agit alors de donner un sens plus concret à un ensemble de caractéristiques locales en les regroupant autour d'un même objet d'intérêt.

II.3.3. Relation de voisinage entre points d'intérêt

La notion de relation entre les caractéristiques locales de l'image est également montrée dans [ChYZ14] et [ToFW08]. Les auteurs exploitent les *mots visuels* pour en construire ce qu'ils appellent des *phrases visuelles*, dans un contexte de reconnaissance de monuments à partir des appareils mobiles. En adoptant le modèle BOW comme représentation d'image pour la reconnaissance d'objets, l'approche proposée construit un voisinage de mots visuels liés les uns aux autres par la méthode des k plus proches voisins, au lieu de considérer les mots visuels de façon indépendante. Cependant, la corrélation entre les caractéristiques visuelles locales existe seulement dans un même voisinage de taille prédéfinie et non pour tous les points clés extraits pour un même objet.

Afin de donner un sens à un objet entier, il serait plus pertinent de déterminer la corrélation entre les caractéristiques locales extraites dudit objet d'intérêt plutôt que d'associer sans réelle signification un point à tous ses voisins.

Une approche légèrement différente, décrite dans le paragraphe suivant, met en avant une recherche plus fine, adapté aux caractéristiques des contours des objets recherchés.

II.3.4. Recherche par représentations de contours

Dans ce cadre, il est proposé par [LiSh02] de prendre en compte et décrire les contours des différents objets apparaissant dans une scène complexe. Les différents contours détectés sont ainsi classifiés selon leur couleur, leur direction et leurs différentes caractéristiques spatiales. Comme il n'est pas possible de définir un seuil global pour la détection seule des différents contours d'objets dans une scène complexe, les auteurs adoptent une approche différente.

Le principe consiste à définir plusieurs critères pour former des clusters cohérents des contours détectés selon les couleurs, leur orientation et leurs caractéristiques spatiales. Ainsi, un

segment de contour est supposé diviser l'espace en deux sous-espaces de couleurs différentes attribuant ainsi une paire de couleurs à un segment donné. Les différents bords détectés peuvent ainsi être regroupés selon leurs couleurs attribués et seuls les plus représentés sont conservés.

Une seconde étape de filtrage est ensuite appliquée, à l'aide d'une technique de classification prenant en compte les orientations des segments détectés. Chaque groupe de segments regroupés par paire de couleurs est ainsi classifié de nouveau dans le but de ne conserver que les bords parallèles (moyennant les différentes perspectives possibles).

De ces différents clusters, les segments conservés sont ensuite projetés sur les deux axes x et y du repère image. Un troisième filtrage est alors effectué, pour regrouper les segments étant proches tant sur l'axe horizontal que sur l'axe vertical. Les contours conservés au final sont les contours persistants de l'image regroupés dans différents clusters représentant qui sont sensés correspondre aux différents objets de l'image. L'approche proposée ici consiste donc à regrouper les lignes isolées précédemment en différents clusters selon leur orientation, leur couleur et leur localisation. Afin de définir quels groupes de segments représentent un bâtiment, le modèle considère un bâtiment comme étant le cluster regroupant le maximum de lignes d'intersection.

Ce modèle est cohérent pour une image d'un bâtiment mis en évidence, cependant, si le monument recherché est situé au second plan avec d'autres bâtiments environnants, il devient plus difficile de repérer l'objet d'intérêt parmi les objets similaires de la scène.

Le problème de l'utilisation des descripteurs locaux comme base de la représentation est soulevé par [LiAl09]. Les auteurs montrent qu'il n'est pas pertinent de considérer seulement des caractéristiques visuelles de bas niveau sans tenir compte de l'information sémantique qui peut être obtenue pour les différents objets. Il s'agit ici de trouver une relation entre les différents descripteurs locaux afin de trouver une connexion entre les différents points ou lignes d'intérêt et d'en déduire de réels objets.

C'est aussi le point de vue apporté par [SiIt07]. Les auteurs proposent d'utiliser le modèle GIST dont le but est de diviser une image déjà traitée selon un modèle de carte de saillance en plusieurs sous-régions, définissant différents objets potentiels. Cependant, cette approche propose une détection de bâtiments à un niveau d'image global. Cela est pertinent lorsque l'objet d'intérêt est prépondérant dans l'image mais n'est plus applicable dès lors que le monument recherché est représenté dans une scène plus complexe, avec des objets occultants.

Une nouvelle famille d'approches propose d'affiner l'information utile exploitée pour décrire une image, afin de se concentrer uniquement sur les éléments pertinents, issus de l'objet d'intérêt au lieu de considérer l'ensemble des points clefs d'une image.

II.3.5. Définition d'une région d'intérêt comme requête

Le travail de [Chen13] montre notamment que lors de la recherche d'un objet spécifique dans une image, il n'est pas efficace d'utiliser l'ensemble de l'image comme requête. Par conséquent, la requête est définie de manière plus fine, comme une région d'intérêt (*Region of Interest* − ROI) dans l'image requête, comme proposé dans [Chen13] et [ArZi12].

Dans ce cadre, il est possible de définir une ROI pour spécifier l'objet recherché uniquement pour les images utilisées en requête. Ce processus est tout à fait possible lors de la spécification des

requêtes par l'utilisateur, à l'aide par exemple d'une interface interactive dédiée. Néanmoins, les images dans le reste de la base de données conservent l'ensemble original des points clés extraits sur l'image entière. Comme nous l'avons vu précédemment, cette information supplémentaire peut interférer avec l'information utile et ainsi influer négativement sur les résultats de la recherche. Sur la base de cette idée de ROI, une solution serait d'appliquer cette méthode sur toutes les images dans la base de données, c'est-à-dire d'isoler uniquement les descripteurs locaux extraits au niveau de l'objet d'intérêt pour toutes les images et non plus uniquement pour l'image requête. Cependant, la question centrale qui se pose est de pouvoir déterminer automatiquement cette ROI. En outre, dans [CMPM11] les auteurs montrent que d'intégrer l'objet recherché en tant que ROI conduit à des résultats plus précis puisque cette approche mets en œuvre uniquement l'information utile de chaque image. La question est alors d'isoler l'objet d'intérêt dans toutes les images de la base de données de façon automatique. C'est bien un objectif central auquel nous nous sommes attaché dans nos travaux.

Dans [BuZa13b], les auteurs proposent une méthode faisant intervenir un masque de sélection de l'objet d'intérêt dans une image représentant une scène complexe. Ce masque binaire isolant la zone d'intérêt est appliqué aux images utilisées en requête de façon à préciser l'objet recherché. Ici encore, les auteurs soulignent l'intérêt de disposer d'une méthode automatique pouvant délimiter les objets d'intérêt candidats à partir de scènes complexes.

Dans [AlPP07], les auteurs s'intéressent à la reconnaissance de bâtiments en passant par la reconnaissance des fenêtres et montrent de façon empirique que la simple détection de fenêtres est suffisante, jusqu'à un certain point, pour reconnaître un bâtiment et définir une région d'intérêt. Dans le cadre spécifique de la reconnaissance de bâtiments, ils proposent une solution fondée sur la détection et la reconnaissance des fenêtres des bâtiments permettant par association de retrouver le bâtiment correspondant. Selon le modèle proposé, une partie seulement des fenêtres d'une façade d'un bâtiment suffisent à reconnaître le bâtiment recherché. Cela implique cependant que les fenêtres non détectées soient similaires à celles retenues ou que les fenêtres détectées soient suffisamment discriminantes pour ledit bâtiment. Cette approche permet de réduire considérablement l'information nécessaire à la reconnaissance d'un bâtiment, mais n'est applicable que dans le cas particulier où la façade du bâtiment recherché est clairement représentée sur l'image et que le monument d'intérêt compte effectivement des fenêtres, ce qui n'est pas toujours le cas pour les monuments touristiques.

II.3.6. Réduction et précision des mots clefs détectés

Dans [TuLo09], les auteurs proposent une méthode pour affiner les caractéristiques des images, afin de conserver seulement l'information utile, dans le cadre d'une méthode par modèle BOW. Le principe consiste à considérer chaque image de la base de données comme une requête. Les images les mieux classés par les résultats de la recherche d'image sont vérifiés géométriquement à l'aide de la méthode du *RANdom SAmple Consensus* (RANSAC). Seulement les points clés géométriquement persistants sont ainsi conservés. Même si cette méthode d'apprentissage n'est pas supervisée, elle comporte deux étapes supplémentaires qui doivent être effectuées de nouveau pour l'ensemble de la base de données pour chaque ajout d'image. Par conséquent, le travail [TuLo09] montre que la réduction du nombre de caractéristiques locales dans la base de données améliore bien les résultats pour la recherche d'images et ceci d'autant plus lorsque les caractéristiques retenus ne représentent que l'objet d'intérêt. Cependant, l'approche reste trop lourde puisque l'ensemble du

processus doit être calculé à nouveau pour toute la base de données à chaque fois que cette base est actualisée, en intégrant de nouvelles images.

Une solution envisageable consiste à segmenter les différents mots visuels extraits d'une image et de ne sélectionner que les mots les plus pertinents décrivant le monument recherché. C'est l'approche proposée par [WeGo09], où un vocabulaire de mots visuels sous forme de patches de taille normalisée est défini. Cependant, la méthode n'est applicable que pour des objets de tailles similaires. En effet, les photographies de monuments peuvent être prises à des distances et zooms différents et donc présenter un même objet de tailles relatives différentes dans différentes images. La méthode proposée par [WeGo09] est appliquée dans le cas d'images satellitaire, où une telle normalisation en taille est naturellement possible mais n'est pas applicable dans le cas de photographies prises par des utilisateurs quelconques.

La problématique du trop grand nombre de descripteurs extraits initialement dans un contexte urbain est aussi soulevé par [LHSA14]. Les auteurs proposent de réduire la dimension des objets ou de les regrouper en grappes similaires.

Le même principe de réduction du nombre de descripteurs est également revisité par [LiAl13], où il est proposé de simplifier le processus d'extraction des descripteurs locaux d'une image. Cependant, en réduisant la dimensionnalité des descripteurs locaux, il existe, en contrepartie, un risque de réduire le potentiel discriminant des mots visuels extraits dans une image.

Une autre approche de filtrage des descripteurs est proposée par [BuZa13b]. Il s'agit ici de négliger les mots visuels répétitifs parmi tous ceux détectés de prime abord, pour ne conserver que les points d'intérêts remarquables dans un unique histogramme de fréquences du modèle BOW. Ici, le contexte d'application concerne l'objectif de retrouver une image dans une séquence vidéo. L'idée proposée par les auteurs est de sélectionner plusieurs images dans la séquence vidéo d'une scène donnée, pour ne retenir que l'image la plus pertinente. Cette pertinence est mesurée en fonction du nombre de points d'intérêts locaux discriminants retenus. Il s'agit donc de filtrer l'ensemble des descripteurs locaux extraits initialement dans une image de façon à ne conserver qu'une information plus pertinente pour la suite du processus. Cette présentation montre bien l'importance de sélectionner l'information utile entre tous les descripteurs locaux détectés *a priori*. Cependant, dans le cadre d'une reconnaissance de bâtiments, des points d'intérêts issus d'un bâtiment recherché peuvent être redondants en fonction de l'architecture du bâtiment. De ce fait, il n'est pas envisageable dans notre cas de les négliger. Le critère de sélection proposé par [BuZa13b] n'est donc pas applicable dans notre situation.

La problématique soulevée par [IBPQ14] concerne les descripteurs locaux extraits dans les régions à faible et fort contraste d'une image. Les auteurs montrent qu'un échantillonnage dense des mots clés détectés sur une image (ou une région de l'image) peut au final ajouter du bruit au lieu d'apporter de l'information. Cette difficulté peut se produire en particulier dans une image représentant une scène complexe, ce qui est le cas dans un environnement urbain.

Dans [SiZi03], les auteurs proposent de supprimer les mots visuels trop répétitifs. Un descripteur local est considéré comme étant à faible contraste et donc répétitif si sa norme est inférieure à un seuil défini de façon empirique. La méthode présentée néglige ces mots visuels et ne les prend pas en compte lors de la construction du vocabulaire dans le modèle de BOW.

Cependant, l'approche introduite dans [IBPQ14] propose d'attribuer à tous les descripteurs dits de faible contraste une valeur par défaut, ne subissant pas le traitement de normalisation usuelle. Une fusion des deux ensembles de descripteurs est effectuée lors de la constitution des histogrammes

du modèle de BOW dans lequel tous les descripteurs considérés comme répétitifs correspondent à l'unique valeur par défaut définie précédemment. Cette approche permet de prendre en considération tous les descripteurs extraits initialement et donc toute l'information disponible dans une image donnée. Cependant, dans un paysage urbain de nombreux bâtiments peuvent être présents sur une image. Dans ce cas, les mots visuels extraits au niveau des bâtiments sont représentés de nombreuses fois et risquent donc d'être négligés, alors que nous cherchons ici à retrouver un monument en particulier.

Dans [TrKJ08], il est proposé de classifier les différentes images d'une base de données en deux groupes : celles contenant un bâtiment et celles n'en contenant pas. La problématique est donc de pouvoir reconnaître un bâtiment dans une scène complexe de façon suffisamment précise pour pouvoir signifier l'absence de bâtiment dans certains cas. Une première étape consiste à détecter les contours des différents objets de l'image. Seuls les segments ayant le même point de fuite sont considérés comme potentiels bâtiments. Les façades des bâtiments peuvent ainsi être représentées par des parallélogrammes. Ceux de même couleur sont ensuite regroupés et considérés comme représentant un même objet. Les descripteurs SIFT issus de ces zones ainsi délimitées sont enfin comparés à ceux de la base de données. Cela permet par la suite de déterminer si un bâtiment est présent ou non dans une image. Cependant, les auteurs montrent que la présence de plusieurs bâtiments en plus de celui recherché pose problème pour la reconnaissance du monument d'intérêt. L'approche proposée permet de préciser si une image représente ou non un bâtiment et ainsi classer une base de données en deux catégories. Toutefois, il n'est pas aisé par de reconnaître un bâtiment d'intérêt, en particulier lorsque l'image le représente dans une scène plus complexe, avec d'autres bâtiments environnants.

Une approche différente est proposée par [ChHH09], où les auteurs utilisent des représentations sous forme d'esquisses. Le principe consiste à mettre en avant les structures prépondérantes d'un bâtiment tels que les fenêtres et les portes, à l'aide d'une méthode de détection des régions d'extrema les plus stables au maximum (*Maximally Stable Extremal Regions detection – MSER*) [EMCU02]. Par la suite, les patchs locaux sont regroupés par la méthode des K-moyennes et comparés entre l'image requête et les images de référence de la base de données. L'avantage de la méthode proposée est une grande robustesse aux variations de points de vue, incluant des prises de vues aériennes. Cependant, la méthode est appliquée uniquement sur des immeubles de bureaux présentant la particularité de montrer des structures déjà discriminantes et redondantes sur l'ensemble du bâtiment. Le contexte d'application de cette méthode reste par conséquent relativement restreint.

II.3.7. Utilisation de machines à vecteurs de support

Dans [LiCH09], les auteurs proposent d'utiliser de classifieurs par machines à vecteurs de support (*Support Vector Machines −SVM*) [Vapn98] pour classer les images dans des base de données à grande échelle. Dans cette approche, les auteurs utilisent un modèle SVM multi-classes pour chaque catégorie spécifiée dans la base de données et comparent les images les unes aux autres. Cette idée de regrouper l'information similaire peut être mise en place de façon analogue au niveau local. En effet, au lieu d'associer une information globale issue d'une image entière et de les comparer image par image, il est possible de construire un classifieur SVM pour distinguer les objets concrets à l'intérieur d'une image et mettre en évidence l'objet d'intérêt parmi le reste de l'image.

Dans [LiCH09], un modèle SVM multi-classe global est entraîné, mais il est également montré dans [CDFW04] qu'un ensemble de modèles SVM binaires peuvent être entraînés pour chaque catégorie dans la base de données. Comme l'objectif est de localiser un bâtiment d'intérêt dans une image, une approche binaire est dans ce contexte plus pertinent. En effet, la problématique principale est ici de trouver automatiquement la région d'intérêt la plus appropriée dans une image de contexte urbain. A partir d'une classification binaire dans lequel la première classe est le bâtiment recherché et la deuxième classe inclus le reste de l'image (les personnes, les véhicules, la végétation, les bâtiments environnants…), il devient possible de construire un ensemble de modèles SVM, avec un classifieur pour chaque catégorie dans la base de données, afin d'isoler le bâtiment recherché sur une image représentant une scène urbaine complexe.

II.3.8. Réseaux neuronaux par apprentissage profond

Récemment, les réseaux neuronaux par apprentissage profond ont montré d'excellents résultats dans le domaine de la reconnaissance visuelle de catégorie d'images comme présenté dans [SiVZ13], [Beza16] et [WZLZ16].

Dans [SiVZ13], il est proposé de mettre en œuvre un réseau neuronal convolutif [LBBH98] dont les différentes couches sont encodées sous la forme d'un vecteur de Fisher [PeSM10]. L'objectif du réseau neuronal est ici de paralléliser le travail de représentation d'une image donnée par portions au travers des différentes couches du réseau pour finalement agréger les différentes régions.

Les auteurs de [Beza16] proposent aussi une implantation d'un réseau neuronal convolutif pour la reconnaissance de bâtiments. Le problème soulevé par les auteurs est celui de la mémoire utilisée. Ils proposent alors une solution de parallélisation sur GPU. Ces aspects constituent une des limites des méthodes par réseaux neuronaux, qui nécessitent de capacités de mémoire et de puissance de calcul importantes et souvent incompatibles avec les ressources des terminaux mobiles grand public aujourd'hui disponibles.

En outre, le principal inconvénient de ces modèles de réseaux neuronaux pour la reconnaissance d'images concerne l'étape d'entraînement du réseau, qui nécessite la disponibilité d'un grand nombre de données annotées, formant une vérité terrain. Cette problématique est soulevée par [WZLZ16], où les auteurs montrent la nécessité de regrouper un grand nombre d'échantillons connus pour entraîner un réseau neuronal donné. Cela requiert notamment un effort non négligeable de la part de l'utilisateur. La solution proposée par les auteurs est alors de n'entraîner qu'une partie des données et d'utiliser les prédictions les plus fiables comme nouvelles données d'entraînement pour la couche suivante du réseau.

Cependant, malgré ces quelques éléments de solution, le problème de la disponibilité d'un nombre suffisamment grand de données d'entraînement persiste. Il est nécessaire de disposer d'au moins une centaine d'images d'entraînement par bâtiment pour entraîner le réseau neuronal alors que nous cherchons à n'utiliser que quelques images dans notre cas. La robustesse du réseau ainsi que sa capacité de généralisation dépendant grandement de son entraînement, il n'est pas toujours fiable de réutiliser des données déjà prédites une première fois en tant que données d'entraînement.

II.4. Bilan

De l'analyse des travaux de l'état de l'art, il ressort que nous pouvons augmenter la précision des résultats de recherche en améliorant l'information utile extraite des images à la fois de la base de données et des images requêtes.

En effet, par analogie avec la recherche textuelle de documents, plus les mots clés utilisés en requête sont précis, plus les documents récupérés en résultat sont pertinents.

Dans cette thèse, nous proposons donc de mettre en œuvre des descripteurs locaux permettant une information plus précise bien que parfois exhaustive. Dans ce cadre, nous avons focalisé nos recherches sur l'élaboration et la mise en œuvre d'une méthode de sélection/filtrage des descripteurs qui vise à conserver uniquement l'information pertinente par rapport à une requête donnée.

Cette approche permet de prime abord de supprimer l'information inconstante apportée par les objets de l'environnement tels que les piétons, les véhicules ou la végétation. La méthode doit permettre de plus d'appréhender la reconnaissance d'un monument particulier dans des images représentants de multiples bâtiments.

III. BASE DE DONNEES D'EXPERIMENTATION

Résumé. Ce chapitre décrit en premier lieu les principales bases de données images publiquement disponibles pour des objectifs de reconnaissance de bâtiments/monuments historiques. Les principales caractéristiques de chacune de ces bases sont présentées, en termes de nombre d'éléments, catégories, vérité terrain ou encore images requêtes. L'analyse de ces éléments nous conduit à retenir, pour nos expérimentations les bases Paris6k et Oxford5k, qui répondent tout à fait à nos objectifs, avec un nombre satisfaisant d'éléments et de catégories représentées. Nous concluons ce chapitre avec une analyse plus approfondie de ces deux bases, en mettant en évidence avantages et limitations de chacune.

Mots clés : bases de données images de bâtiments, vérité terrain, catégories d'images, bases Paris6k et Oxford5k.

Quelles que soient les techniques de recherche et de reconnaissance d'images utilisées, il est nécessaire de disposer d'une base de données d'images définie et sémantiquement catégorisée en classes pertinentes et suffisamment représentées.

Ce cadre système permet entre autre de vérifier et évaluer les résultats de recherche en les quantifiant par rapport à une vérité terrain connue.

Nous présentons dans ce chapitre les bases de données d'images utilisables dans ce contexte et prise en compte dans la littérature ainsi que les celles retenues pour nos exemples et expérimentations dans le cadre spécifique des images urbaines.

Dans le but de créer un cadre défini dans lequel il est possible d'évaluer les performances des différentes méthodes de recherche d'image, nous utilisons des bases de données de références regroupant des monuments touristiques.

III.1. Bases de données de bâtiments

Commençons par présenter les différentes bases de données d'images recensées dans la littérature et utilisées pour la reconnaissance d'images en général.

III.1.1. Base Holidays

La base de données Holidays [JeDS08] inclut 500 groupes d'images et 1491 images au total distribuées comme suit. Parmi ces photos, 500 images sont exploitées en tant que requête et 991 images pertinentes correspondant dans la vérité terrain.

Elle est mise à l'œuvre dans les expérimentations des travaux [ArZi13], [JDSP10], [SiZi03] et [PLSP10]. Les photos de cette base de données présentent l'avantage d'avoir été prises intentionnellement de façon à tester la robustesse aux variations de rotations, points de vue et éclairages. Cependant, les images sont d'une trop grande variété et nous pouvons y trouver tant des monuments que des paysages, des feux d'artifices, comme illustré **Figure 19**.

Figure 19. Exemples d'images de la base de données Holidays.

III.1.2. Base Flickr60k

La base de données Flickr60k [JeDS08] compte 67 714 images issues du réseau social Flickr contenant entre autre une partie de la base de données Holidays décrite précédemment et est utilisée en particulier dans [ArZi13]. De même que la base de données Holidays, il s'agit d'images de divers lieux autour du monde et pas seulement des images de bâtiments.

III.1.3. Base ZuBuD

La base de données ZuBuD [ShSG03] est une base de donnée d'images de bâtiments comptant 1 005 images représentant 201 bâtiments de Zurich (**Figure 20**). Chaque bâtiment est représenté par 5 images différentes. Elle est utilisée notamment dans [ObMa05].

Cependant, les images recensées ici, bien que représentant toutes des bâtiments, ne comptent que très peu d'images différentes d'un même bâtiment, en privilégiant la diversité des monuments photographiés.

Il est plus aisé dans notre cas de prendre en compte un nombre plus réduit de bâtiments divers mais comptant plusieurs images différentes du même dit bâtiment.

Figure 20. Exemples d'images de la base de données ZuBuD.

III.1.4. Base Cityscapes Dataset

La base de données Cityscapes Dataset [CORS15], [CORR16] regroupe 25 000 images annotées en 30 classes issues de 50 villes différentes à différentes saisons et heures de la journée et annotées avec différents tag des conditions météorologiques et GPS.

Cependant, cette base de données n'a été disponible qu'au courant de l'année 2015 et la taille des données proposées par cette base de données ont finalement été trop importantes pour pouvoir être prise en charge lors de nos expérimentations.

D'autre part, la base de données propose des annotations d'images (**Figure 21**) qui sont cependant orientées fortement pour un découpage en objets spécifiques au sein de la scène et donc moins adaptés pour nos objectifs.

Figure 21. Exemples d'images annotées de la base de données Cityscapes Dataset.

III.1.5. Base Paris6k

La base de données Paris6k [PCIS08] compte 6 412 images de Flickr des différents monuments et bâtiments de Paris (**Figure 22**). Elle est utilisée notamment dans les travaux de [ArZi13], [BuZa13a], [BuZa13b], [DGJP13] et [ArZi12]. Cette base de données correspond tout à fait au cadre de notre recherche d'images en milieu urbain.

Figure 22. Exemples d'images de la base de données Paris6k.

III.1.6. Base Oxford5k

La base de données Oxford5k [PCIS07] prend appui sur la structure de la base de donnée Paris6k décrite précédemment. Elle regroupe 5 062 images sur 11 différents lieux d'Oxford illustrés **Figure 23**. Elle est notamment utilisée dans [ArZi13], [PLSP10] et [JDSP10].

Figure 23. Exemples d'images de la base de données Oxford5k.

III.2. Bases de données retenues

Afin de prouver la robustesse des méthodes proposées dans nos travaux, nous conduisons les expérimentations sur des bases de données les plus adaptées, représentant uniquement des monuments en catégories définies et issues de deux différentes villes : Paris6k [PCIS08] et Oxford5k [PCIS07].

III.2.1. Contenu et classes de bâtiments

La base de données Paris6k compte 6 412 images issues de 11 monuments touristiques différents à Paris :

1. L'arche de la Défense,
2. La Tour Eiffel,
3. L'Hôtel des Invalides,
4. Le Louvre,
5. Le Moulin Rouge,
6. Le Musée d'Orsay,
7. Notre Dame,
8. Le Panthéon,
9. Le musée Pompidou,
10. Le Sacré Cœur et
11. L'Arc de Triomphe.

L'ensemble de données Oxford5k compte 5 063 images représentant 11 sites d'intérêt touristique à Oxford :

1. All Souls,
2. Ashmolean,
3. Balliol,
4. Bodleian,
5. Christ Church,
6. Cornmarket,
7. Hertford,
8. Keble,
9. Magdalen,
10. Pitt Rivers et
11. Radcliffe Camera.

III.2.2. Vérités terrains

Dans la base de données Paris6k, pour chaque catégorie de bâtiment, 5 images sont définies en tant qu'image requête et sont disponibles dans la vérité terrain. Au total, 55 images de la base de données sont donc utilisées comme images requête.

Pour chaque catégorie de bâtiment, une partie des images est comptabilisée comme correctement détectée et une autre partie comme erronée. Cela permet par la suite, lors de la phase de recherche d'images, de vérifier que les résultats obtenus pour les images requêtes définies dans la vérité terrain sont au plus des images présentées comme correctement détectées et regroupent un minimum d'images erronées (**Figure 24**).

Figure 24. Exemples d'images de la vérité terrain utilisées en tant que requête pour 3 catégories de la base de données Paris6k. La première ligne correspond aux images de la Tour Eiffel, la deuxième ligne correspond aux images de l'Arc de Triomphe et la troisième ligne correspond aux images de la cathédrale Notre-Dame de Paris.

Au total, la vérité terrain compte 1 791 images correctement détectées, toutes catégories de bâtiments confondues et 1 649 erronées réparties suivant chaque catégorie, comme présenté dans le **Tableau 1**.

La vérité terrain de la base de données Oxford5k est construite de façon similaire à la vérité terrain de la base de données Paris6k (**Figure 25**). Pour chaque catégorie de bâtiment, 5 images sont définies comme image requête, 568 images correctement détectées et 299 erronées réparties suivant chaque catégories comme présenté dans le **Tableau 2**.

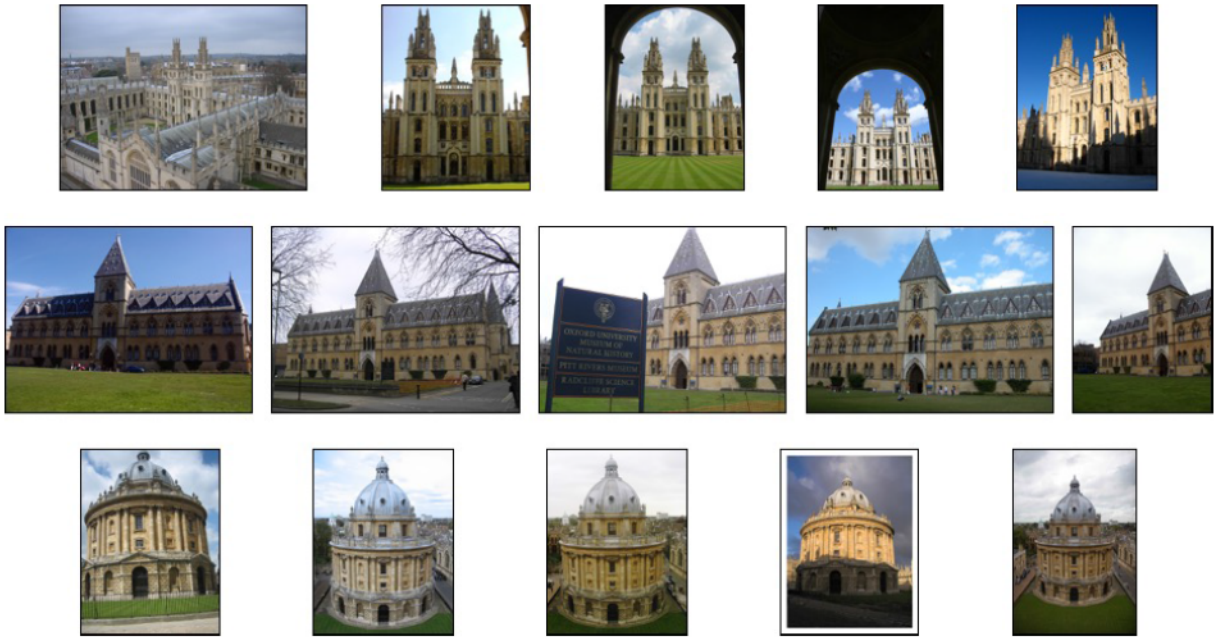


Figure 25. Exemples d'images de la vérité terrain utilisées en tant que requête pour 3 catégories de la base de données Oxford5k. La première ligne correspond aux images de l'Université All Souls, la deuxième ligne correspond aux images du musée Pitt Rivers et la troisième ligne correspond aux images de la bibliothèque Radcliffe.

Catégories	Correctes	Erronées
La Défense	117	116
Tour Eiffel	289	473
Invalides	198	99
Louvre	152	147
Moulin Rouge	237	103
Musée d'Orsay	72	15
Notre Dame	119	112
Panthéon	126	191
Pompidou	51	77
Sacré Cœur	149	152
Arc de Triomphe	281	164
Total	**1791**	**1649**

Tableau 1. Nombre d'images décrites dans la vérité terrain pour chaque catégorie de bâtiment de la base de données Paris6k

Catégories	Correctes	Erronées
All Souls	78	33
Ashmolean	25	6
Balliol	12	6
Bodleian	24	6
Christ Church	78	55
Cornmarket	9	4
Hertford	54	7
Keble	7	4
Magdalen	54	49
Pitt Rivers	6	2
Radcliffe Camera	221	127
Total	**568**	**299**

Tableau 2. Nombre d'images décrites dans la vérité terrain pour chaque catégorie de bâtiment de la base de données Oxford5k

III.2.3. Descripteurs extraits

Les descripteurs extraits sont les descripteurs SIFT des points clefs détectés par la méthode de détection des régions hessienne-affines [MiSc02].

Le nombre de descripteurs extraits varie beaucoup en fonction de l'image prise en compte. Dans la base de données de Paris6k, le nombre de descripteurs extraits varie de 0 à 24 669 points clés détectés avec en moyenne 2 671 points clés détectés sur les 6 412 images.

Dans la base de données Oxford5k, le nombre de descripteurs extraits varie de 0 à 22 540 points clés détectés avec en moyenne 2 807 points clés détectés sur les 5 063 images.

Le nombre de descripteurs par catégorie de bâtiment dans la vérité terrain sont présentés dans le **Tableau 3**. Ici, nous avons considéré un détecteur de type Hessien-affine. Il s'agit des descripteurs extraits expérimentalement suivant la méthode présentée précédemment. Les descripteurs ainsi extraits sont illustrés **Figure 26**.

Paris6k	
Catégories	**Descripteurs**
La Défense	160 584
Tour Eiffel	575 040
Invalides	494 805
Louvre	381 166
Moulin Rouge	584 241
Musée d'Orsay	317 822
Notre Dame	419 977
Panthéon	231 708
Pompidou	242 564
Sacré Cœur	407 761
Arc de Triomphe	712 255
Total	**4 527 923**

Oxford5k	
Catégories	**Descripteurs**
All Souls	258 742
Ashmolean	50 469
Balliol	37 502
Bodleian	71 519
Christ Church	157 298
Cornmarket	32 944
Hertford	161 559
Keble	23 135
Magdalen	134 066
Pitt Rivers	20 012
Radcliffe Camera	683 548
Total	**1 630 794**

Tableau 3. Nombre de descripteurs extraits pour chaque image correcte de la vérité terrain par catégorie de bâtiment dans les bases de données Paris6k et Oxford5k.

Figure 26. Exemples de descripteurs SIFT extraits en utilisant la méthode de détection hessienne-affine pour des images de la base de données Paris6k.

III.2.4. Analyse critique

Bien que ces images correspondent tout à fait à notre cadre de recherche d'images en milieu urbain, la pertinence des résultats obtenus dépend de la vérité terrain fournie dans [PCIS08] et [PCIS07] respectivement pour les bases de données Paris6k et Oxford5k.

Pour la base de données Paris6k, il n'est pas rare de trouver parmi les images considérées comme correctes d'après la vérité terrain des photos ne contenant pas l'objet défini dans les images considérées comme requête. Le monument recherché est parfois représenté sur une photo aérienne

décrivant une zone large de bâtiments, il peut être occulté partiellement ou en grande partie sur une autre, comme illustré **Figure 27** pour l'exemple des images représentant les Invalides dans la base de données Paris6k avec les images considérées en requêtes présentées **Figure 28**. Ce problème est aussi illustré pour la catégorie du Musée d'Orsay **Figure 29** et **Figure 30** mais n'apparaît pas dans la base de données d'Oxford5k qui est plus homogène entre les images requêtes et les images retournées.

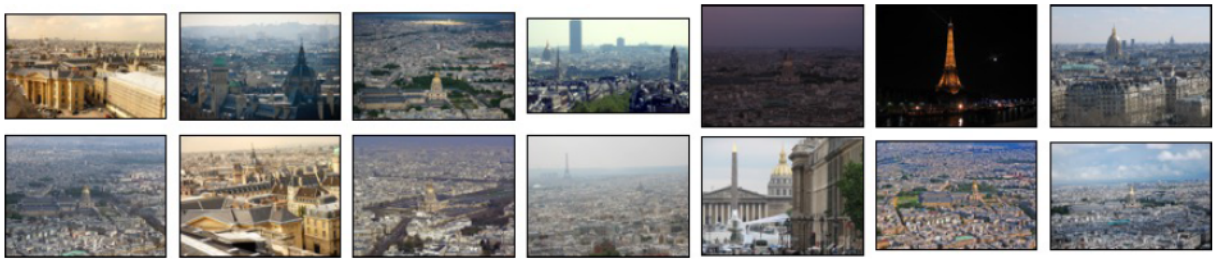


Figure 27. Exemples d'images de la vérité terrain de la base de données Paris6k considérées comme correctement retournées pour la catégorie de bâtiment des Invalides.

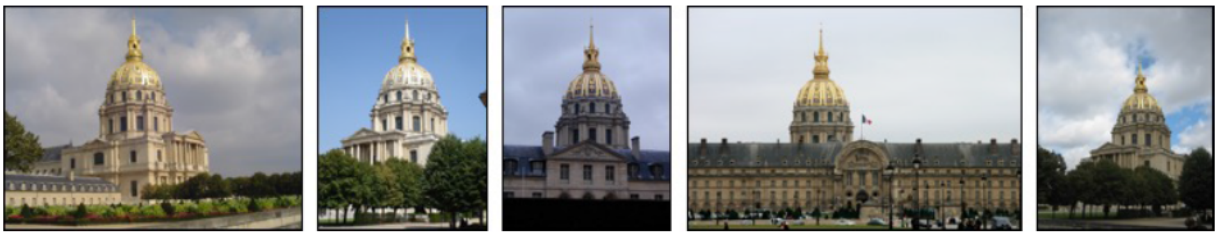


Figure 28. Images de la base de données Paris6k utilisées en tant qu'images requêtes pour la recherche d'images dans la catégorie de bâtiment Invalides.

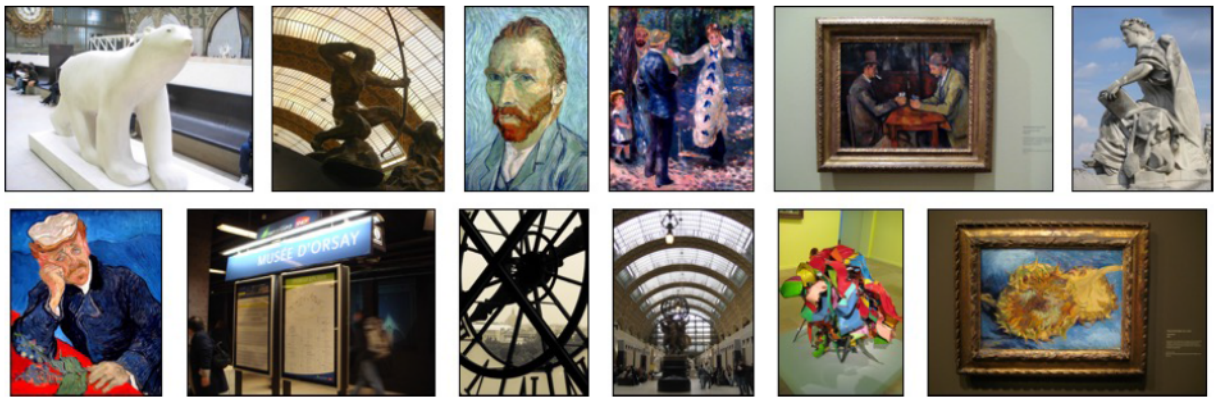


Figure 29. Exemples d'images associées à la catégorie de bâtiment du Musée d'Orsay de la base de données Paris6k.

Figure 30. Images de la base de données Paris6k utilisées en tant qu'images requêtes pour la recherche d'images dans la catégorie de bâtiment du Musée d'Orsay.

IV. CLASSIFICATION GLOBALE DES DESCRIPTEURS LOCAUX

Résumé. Dans ce chapitre, nous introduisons une première méthode de classification globale qui vise, à partir d'une base de données images munie d'une vérité terrain, à réaliser un premier filtrage des descripteurs locaux pour en retenir uniquement ceux correspondant à des régions de bâtiments. Le contexte méthodologique adopté s'appuie sur de descripteurs SIFT associés aux points d'intérêt extraits des images. Quant à la méthode de classification, un formalisme par machine à vecteurs support (SVM) est retenu. Trois stratégies différentes de classification sont considérées, testées et discutées. Elles concernent notamment la prise en compte ou non des éléments de contexte persistant. Le choix des différents paramètres impliqués dans le modèle de classification SVM est également explicité et argumenté. Les résultats obtenus sur la base Paris6k mettent en évidence la supériorité d'une procédure restrictive, qui consiste à isoler au maximum les bâtiments recherchés par rapport aux autres éléments, incluant à la fois des régions persistantes (*i.e.*, bâtiments voisins) et non-persistantes (piétons, véhicules, végétation…).

Mots clés : classification de données, modèle SVM binaire.

Dans ce chapitre, l'objectif est d'investiguer en quelle mesure il est possible d'analyser et d'affiner les descripteurs locaux extraits des images de la base de données pour en retenir uniquement ceux correspondant à la stricte information utile pour représenter les bâtiments et monuments recherchés parmi les catégories de bâtiments définies.

Le cadre méthodologique que nous avons adopté est celui de représentations par points d'intérêt, décrit au **Paragraphe III.2.3**. La détection de ces points est réalisée avec la méthode hessienne-affine [MiSc02]. Le voisinage local de chaque point est ensuite caractérisé à l'aide des descripteurs SIFT. Enfin, pour globaliser l'information, un modèle de type BOW est adopté. Les histogrammes de mots visuels sont construits sur un vocabulaire obtenu avec la méthode k-moyennes.

Cependant, pour une image donnée, des points d'intérêts sont détectés sur l'ensemble de la scène, que ce soit des bâtiments voisins, des piétons, des véhicules, des signes, du mobilier urbain, ou encore de la végétation. Comme illustré **Figure 31**, seule une partie de ces points correspond à l'information réelle extraite du bâtiment d'intérêt tandis qu'une autre partie des points est extraite de l'environnement et n'apporte donc pas d'information pertinente concernant le bâtiment en soi.

Figure 31. Points d'intérêts extraits sur une image : **a)** ensemble original des points extraits, **b)** ensemble des points extraits en particulier du bâtiment d'intérêt, **c)** ensemble des points extraits de l'image mais ne représentant pas le bâtiment d'intérêt.

Notons que le nombre de points d'intérêt qui ne correspondent pas au bâtiment peut être relativement important et même supérieur à ceux utiles, associés au bâtiment recherché.

Nous proposons donc de filtrer cette information de façon à ne conserver que les points d'intérêts extraits du bâtiment recherché et ainsi ne prendre en compte que l'information pertinente. Ce filtrage s'appuie sur une classification SVM et est mis en œuvre tant pour l'ensemble de la base de données pour la partie hors ligne de construction du vocabulaire que lors de la recherche effective, comme illustré **Figure 32**.

Cette approche permet de construire un vocabulaire ne regroupant que les mots visuels les plus discriminants et représentant uniquement les monuments d'intérêt.

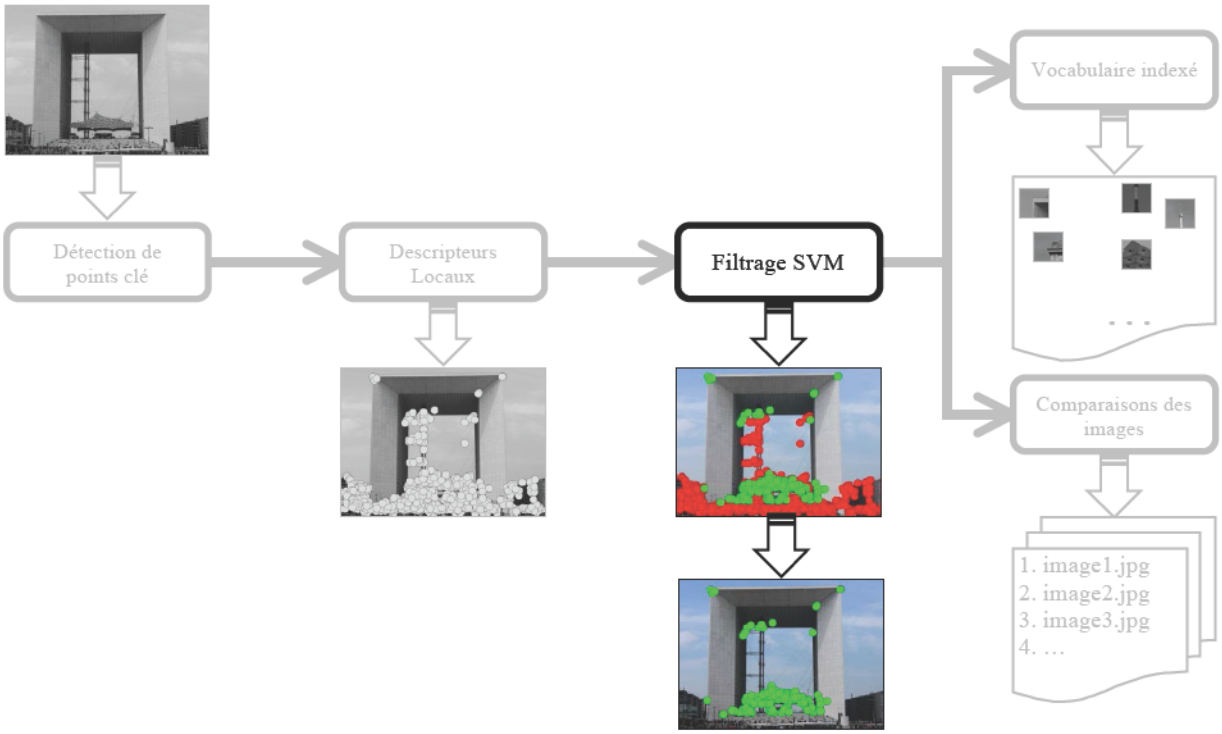


Figure 32. Processus de construction de vocabulaire et de recherche d'image aggrémenté d'une étape de filtrage des points d'intérêts.

Nous cherchons donc à classer l'ensemble des points d'intérêts détectés dans deux classes suivant que le point soit extrait du monument recherché ou non. Comme présenté dans l'état de l'art, la méthode des machines à vecteur de support (SVM) est une technique parfaitement adaptée pour réaliser une classification binaire. Dans ce contexte, les deux classes définies sont *bâtiment* et *non-bâtiment*, illustré **Figure 33**. Chaque descripteur d'une image est donc attribué à l'une de ces deux classes pour n'en retenir que les descripteurs du bâtiment recherché.

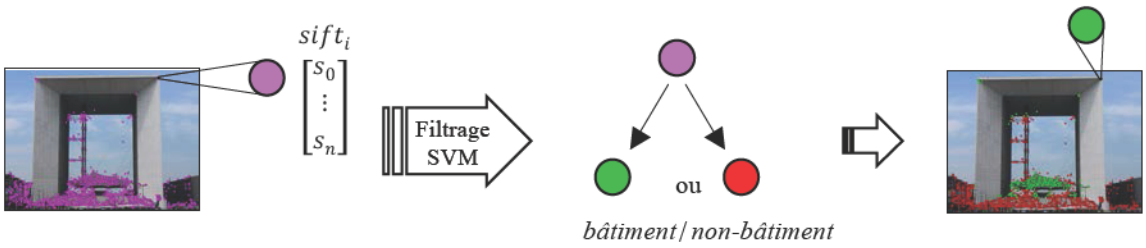


Figure 33. Illustration de filtrage des descripteurs SIFT d'une image en deux classes *bâtiment* et *non-bâtiment*.

Soulignons que l'espace de classification est bien celui des descripteurs locaux SIFT associés à chaque point d'intérêt. Le choix des descripteurs SIFT se justifie par les propriétés d'invariance par rapport à la position et l'angle de vue. Ainsi, un même monument peut être représenté sur une photo dans différentes positions et de plus ou moins loin, comme illustré **Figure 34**, ce qui ne permet pas de généraliser aisément une classification.

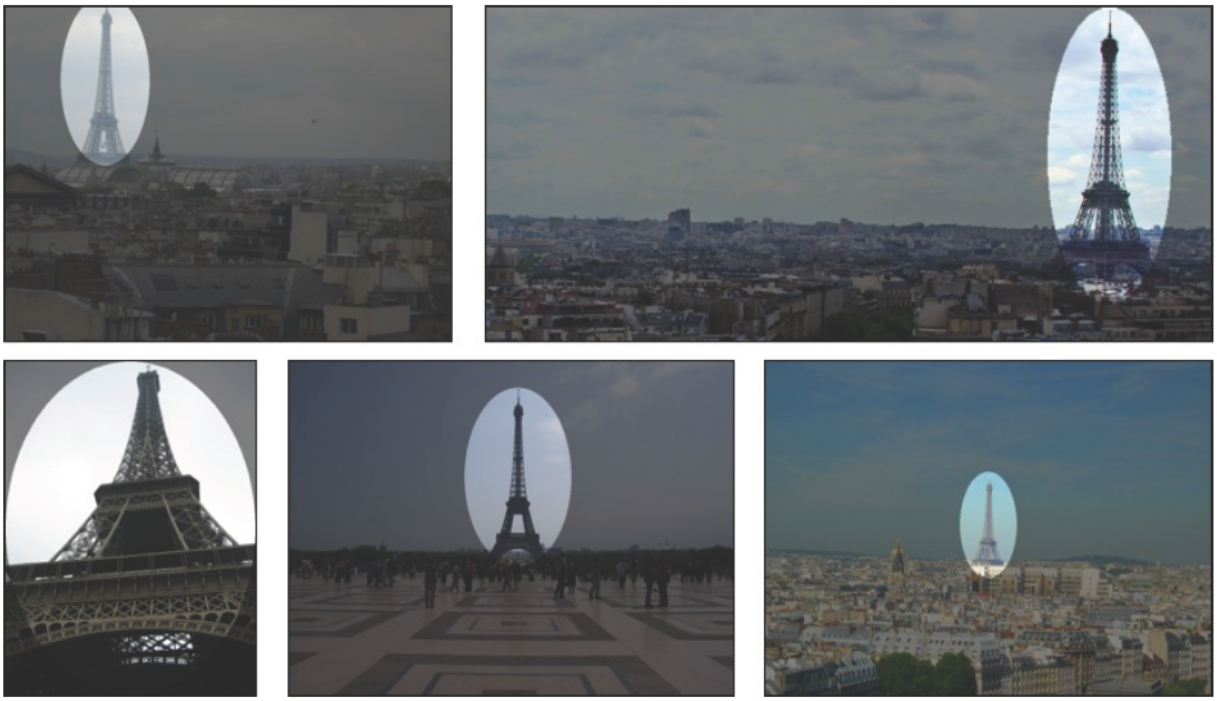


Figure 34. Le bâtiment d'intérêt peut être à différentes positions dans une image et de tailles différentes. Exemple d'images de la Tour Eiffel dans la base de données Paris6k.

Rappelons à présent le formalisme de classification SVM adopté.

IV.1. Classification par machines à vecteurs de support (SVM)

Le modèle SVM (*Support Vector Machines*) est une méthode de classification supervisée présentée dans [BoGV92], qui a été utilisée extensivement ces dernières années à des objectifs de classification d'images [ArZi12], [LHSA14], [Chen13], [LiCH09], [LiAl13]. Rappelons-en le principe, dans un contexte de classification binaire.

A partir d'un ensemble de données d'apprentissage séparées en deux classes, l'objectif est de déterminer une frontière de séparation entre ces deux classes comme illustré **Figure 35** pour un exemple 2D.

Si l'on considère une séparation linéaire de l'espace, la fonction de décision est de la forme $D(x) = \text{signe}(f(x))$ avec $f(x) = ax + b$ pour tout vecteur x de l'espace et a et b deux paramètres classifiant au mieux les données d'entraînement, c'est-à-dire tel que $D(x_i) = y_i$ pour tous les points d'entraînement (x_i, y_i).

On associe à cette fonction de décision une frontière de décision définie selon l'équation (20) :

$$\Delta(a, b) = \{x \in \mathbb{R} \mid ax + b = 0\} \tag{20}$$

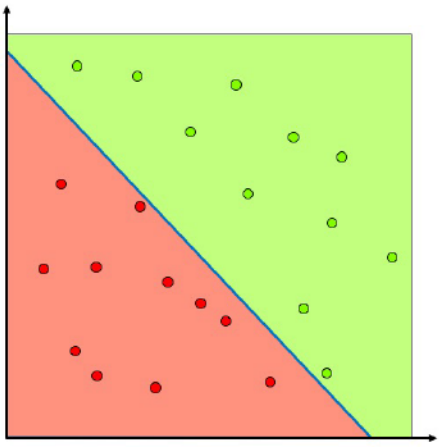


Figure 35. Exemple de classification SVM en deux dimensions.

L'entraînement du modèle SVM revient donc à un problème d'optimisation formalisé par l'équation (21) pour tous les points (x_i, y_i) des données d'entraînement. On introduit dans la suite un paramètre $e_i \geq 0$ définissant le rapport de distance du point i à la frontière de séparation tel que si $e_i > 1$ le point est considéré comme étant mal classifié. Nous définissons ainsi une marge "souple" autour de la frontière de séparation, comme illustré **Figure 36**.

$$\min_a \|a\|^2 \quad \text{tel que } y_i(ax_i + b) \geq 1 \tag{21}$$

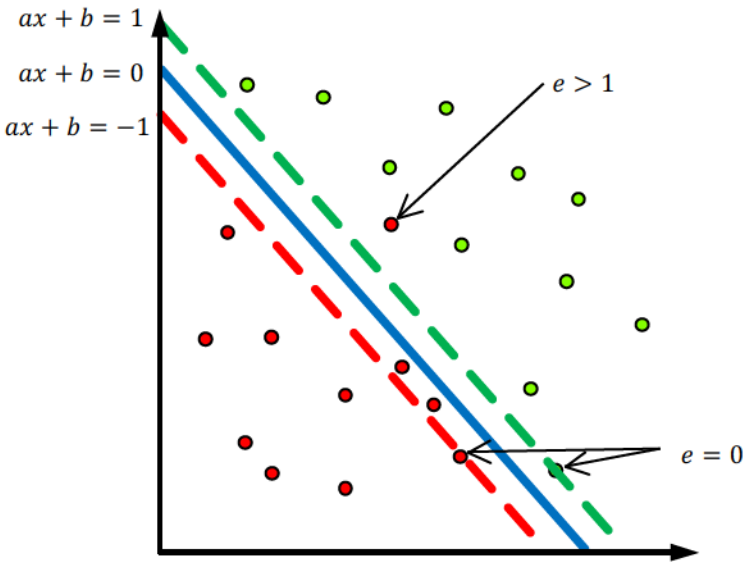


Figure 36. Définition d'une marge dite "souple" dans un modèle d'entraînement SVM.

L'équation d'optimisation devient alors (22) pour tous les points (x_i, y_i) des données d'entraînement. Nous introduisons dans cette équation le paramètre C, appelé paramètre de régularisation. Plus ce paramètre C est petit, plus les contraintes sont souples et donc plus la marge de séparation est large. À l'inverse, plus la valeur de C est grande, plus la marge est étroite.

$$\min_a \|a\|^2 + C \sum_{i=1}^{N} e_i \text{ tel que } y_i(ax_i + b) \geq 1 - e_i \tag{22}$$

En s'appuyant sur la bibliothèque logicielle présentée dans [ChLi11], le modèle le plus adéquate est le *C-Support Vector Classification* (C-SVC). Ce modèle permet une séparation imparfaite des classes avec un paramètre C pénalisant les valeurs mal classées. Le paramètre C représente dans ce cas le compromis entre la tolérance de classification erronée et la simplicité de la limite de décision. Plus la valeur de C est faible, plus l'hyperplan de séparation est lisse mais aussi imparfait. A l'inverse, plus la valeur de C est élevée, plus la classification est précise mais aussi plus longue à mettre en place.

Le modèle de classification SVM calcule un hyperplan partitionnant l'espace en deux sous-espaces. Cette limite dépend du noyau SVM choisi parmi les modèles linéaire, polynômial ou radial. La fonction linéaire étant trop simple, nous ne conservons que les noyaux polynômiaux et radiaux. Les

équations de marge de ces deux noyaux sont rappelées dans les équations (23) et (24) pour deux vecteurs u et v de l'espace des caractéristiques à classifier.

$$k(u,v) = (\gamma\, u' \times v + r)^{\delta} \tag{23}$$

$$k(u,v) = \exp(-\gamma\, |u - v|^2) \tag{24}$$

Dans la première équation polynomiale, r est un terme d'interception utile pour mettre à l'échelle les données et ajouter un éventuel scalaire de compensation. Pour le noyau polynômial, un paramètre important à définir est le degré du polynôme δ. Un polynôme de degré supérieur fournit logiquement un classifieur plus précis au détriment de la complexité de calcul et du temps de traitement.

Dans les deux cas, des noyaux des SVM polynomial et radial, le paramètre γ représente l'influence d'un vecteur d'entraînement. Plus la valeur de γ est élevée, plus l'influence de chaque donnée a de l'importance et plus la marge de décision se rapproche de chacune des données d'apprentissage. Avec un grand nombre de données, ce paramètre γ influence grandement l'équation de l'hyperplan puisqu'il est attribué à chaque point clé un poids plus ou moins important d'attraction de la frontière.

En effet, si la valeur de γ est trop élevée ou trop basse, le modèle SVM classe toutes les données dans une des deux classes seulement. Plus précisément, si γ est trop élevé, l'influence d'un vecteur lors de la phase d'entraînement sur son voisinage est en fait réduite au vecteur lui-même. Par conséquent, seul ce vecteur peut être classé dans cette classe et aucune autre donnée en entrée ne sera attribuée à cette classe, aussi proche soit elle du vecteur d'entraînement. Cela correspond à un problème de sur-ajustement *(overfitting*). Dans ce cas, seules les données exactes utilisées en entraînement peuvent être correctement classées et aucune flexibilité n'est permise. D'autre part, si γ est trop faible, le voisinage d'un vecteur d'entraînement regroupe en réalité toutes les données d'entraînement et donc toutes les données sont classées dans cette unique classe. Cette situation correspond à un entraînement trop laxiste acceptant toutes les données dans la première classe définie dans l'entraînement.

Nous retenons au final un modèle polynômial de degré 5. Pour les polynômes de degré supérieur (c'est-à-dire de degré supérieur ou égal à 6), les résultats de la classification SVM sont équivalents, mais les temps de calcul pour l'entraînement et la prédiction augmentent considérablement.

Après plusieurs expérimentations, consistant à faire évoluer la valeur du paramètre C, il apparaît que $C = 10$ correspond à un compromis satisfaisant entre l'acceptation de l'imperfection et de la complexité de calcul. Le terme r d'interception n'étant pas nécessaire pour nos données, ce paramètre est défini dans notre contexte à sa valeur par défaut $r = 0$. Après plusieurs expérimentations faisant varier le paramètre γ, la valeur de $\gamma = 1$ présente le meilleur équilibre sur l'influence des deux classes.

Le choix de ces différents paramètres des modèles SVM est argumenté en détails dans la **Section IV.3**.

Quant aux différentes stratégies nécessaires pour la définition des données d'apprentissage, elles sont décrites dans le paragraphe suivant.

IV.2. Classification des descripteurs d'image *bâtiment* et *non-bâtiment*

Le but de cette méthode est d'attribuer à chaque descripteur local extrait d'une image une classe *bâtiment* ou *non-bâtiment*. Nous considérons la classe *bâtiment* comme représentant la classe positive et la classe *non-bâtiment* comme étant la classe négative. D'après cette définition, un point pourra être attribué à l'une ou l'autre classe suivant quatre cas illustrés **Figure 37**. Ici et par la suite, pour l'ensemble de nos illustrations, les points attribués à la classe *bâtiment* sont colorés en vert et ceux attribués à la classe *non-bâtiment* sont colorés en rouge.

Les différents cas pouvant alors se produire sont les suivants. Dans le premier cas, le point étudié est effectivement extrait du bâtiment recherché et est correctement attribué à la classe *bâtiment*. Il s'agit d'un *vrai positif* obtenu par le classifieur. Un deuxième cas pouvant survenir est de retrouver un point attribué de la même manière à la classe *bâtiment* mais n'en faisant en réalité pas parti. Il s'agit d'un *faux positif*, le point ayant été attribué à la mauvaise classe. Cette situation est la plus délicate pour la suite du processus. En effet, pour la construction de notre vocabulaire, nous ne retenons que les points issus de la classe *bâtiment*. Un descripteur ayant été extrait d'une autre partie de l'image que du monument recherché sera donc pris en compte par la suite et pourra ajouter de l'information non-pertinente. Une autre classification pouvant survenir est de correctement attribuer un point ne faisant pas parti de l'image à la classe *non-bâtiment*. Il s'agit d'un *vrai négatif*, correctement classifié. Enfin, la dernière possibilité restante est donc de mal classifier un point faisant en réalité parti du monument recherché et l'attribuer à la classe *non-bâtiment*. Dans ce cas d'un *faux négatif*, le point considéré ne sera tout simplement pas retenu pour la construction du vocabulaire. Cependant, le vocabulaire étant construit à partir des milliers d'images de la base de données, ce manque d'information n'est pas préoccupant pour la suite du développement puisque compensé par l'apport des autres mots similaires.

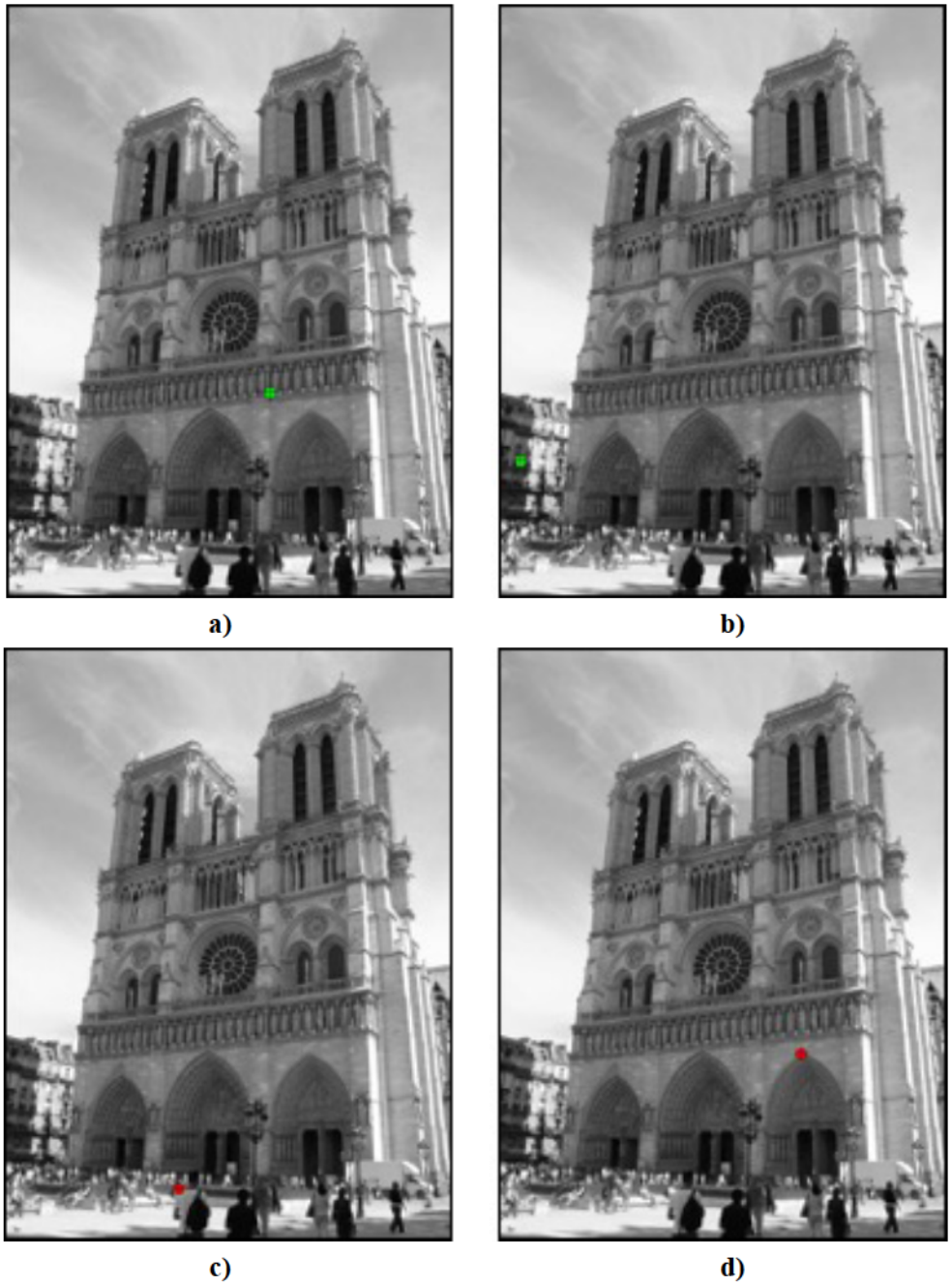


Figure 37. Différents résultats possibles de la classification de points clés dans un exemple issu de la catégorie d'images représentant Notre Dame de la base de données Paris6k **a)** vrai positif **b)** faux positif **c)** vrai négatif **d)** faux négatif.

La première phase du modèle de classification supervisée retenu est donc une étape d'entraînement présentée dans le paragraphe suivant.

IV.2.1. Définition des données d'entraînement

Nous avons décidé d'utiliser en entraînement uniquement les images définies en tant qu'images requêtes dans la vérité terrain. Pour la base Paris6K, nous disposons donc de 55 images pour entraîner un modèle SVM permettant de ne retenir que les descripteurs issus des 11 bâtiments recherchés. Le modèle SVM est ici défini dans l'espace des descripteurs locaux. La dimension de cet espace est donc dans notre cas celle des descripteurs SIFT, c'est à dire de dimension 128.

Le modèle SVM étant une méthode de classification supervisée, il est nécessaire de constituer une vérité terrain. Cela requiert de segmenter manuellement les 55 images requête retenues. Pour cela, nous avons créé un masque binaire pour chaque image, comme illustré sur la **Figure 38**, permettant d'isoler les descripteurs des bâtiments d'intérêts de leur environnement pour l'ensemble de ces 55 images.

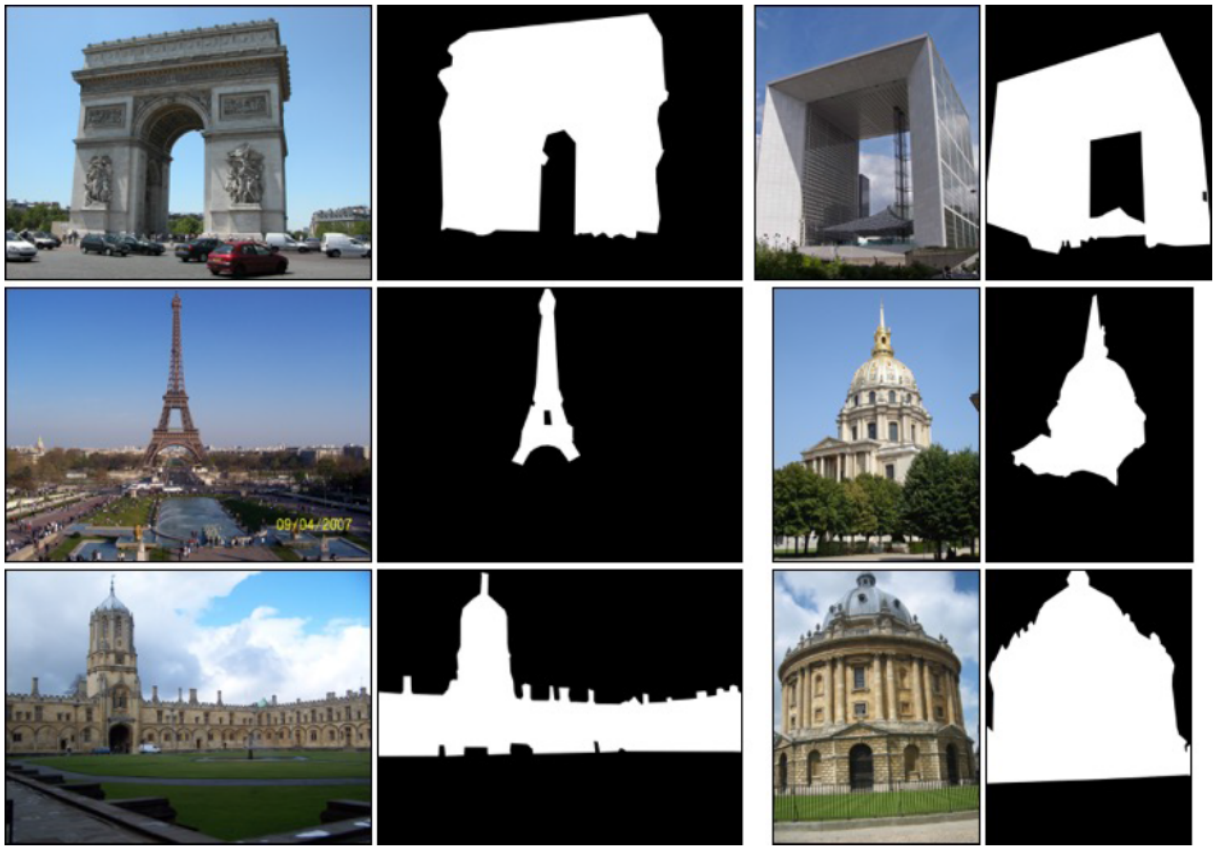


Figure 38. Exemples de masques binaires créés pour l'entraînement des classifieurs SVM pour isoler le bâtiment d'intérêt au sein d'une scène urbaine complexe.

Différentes stratégies d'entraînement sont alors possibles suivant la définition des masques d'entraînement, en fonction des éléments considérés comme faisant partie du bâtiment d'intérêt recherche ou non et avec ou sans prise en compte de son environnement. Nous présentons dans les sections suivantes les différentes approches retenues et testées.

IV.2.1.1. Isolation de l'objet d'intérêt sans prise en compte de l'environnement persistant

Une première approche consiste à considérer que l'ensemble des points d'intérêts extraits de l'image ne forme pas une partition complète de l'espace. En effet, certains points appartiennent clairement à l'une ou l'autre classe mais la question se pose pour une partie des descripteurs. Par exemple, les points issus du bâtiment recherché sont entraînés de façon évidente dans la classe *bâtiment*, alors que ceux issus des piétons, de la végétation ou des véhicules sont attribués de façon supervisée à la classe *non-bâtiment*. Cependant, les autres bâtiments voisins du bâtiment recherché ne peuvent être clairement attribués à l'une ou l'autre classe. En effet, dans un sens ils ne font pas partie du monument mais d'un autre point de vue, ils pourraient conduire à exclure des points appartenant à une autre catégorie de bâtiment et ainsi réduire l'information conservée pour la suite du processus. Ces points particuliers ne sont donc pas pris en considération dans aucune des deux classes dans le but de ne favoriser ni la classe *bâtiment* ni la classe *non-bâtiment*.

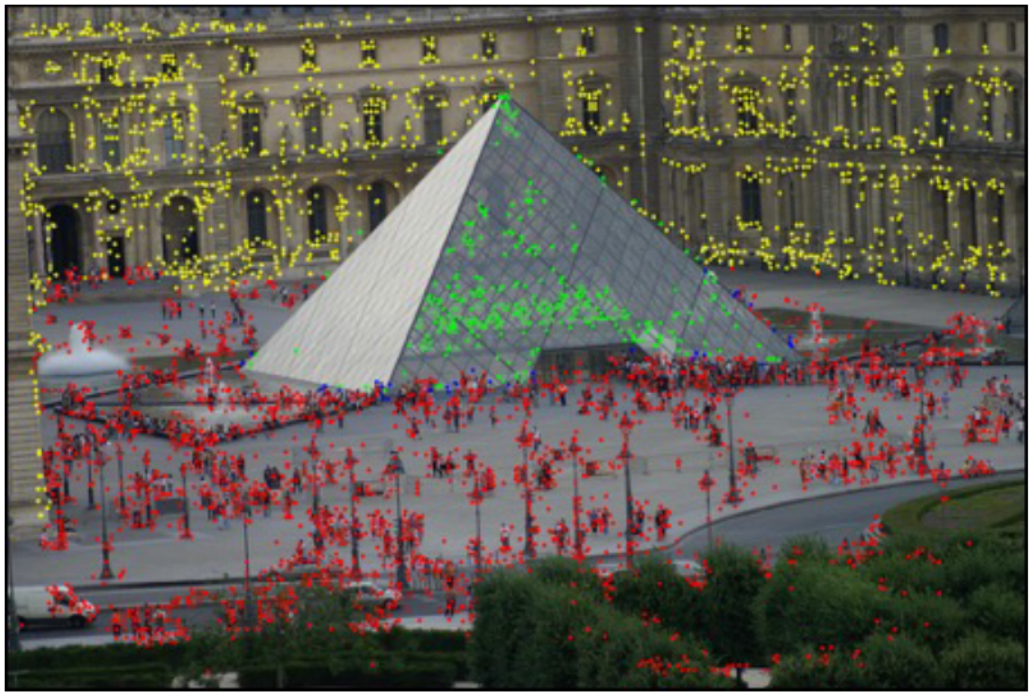
Nous proposons alors de définir deux masques différents pour isoler les points d'entraînement de la classe *bâtiment* et ceux de la classe *non-bâtiment*. Il est à noter que ces masques ne sont donc pas nécessairement complémentaires. Sur les exemples illustrés **Figure 39** et **Figure 40** les points verts représentent les descripteurs associés à la classe *bâtiment* de façon catégorique pour l'entraînement et les points rouges ceux associés à la classe *non-bâtiment*. Les points jaunes correspondent aux points appartenant aux bâtiments environnants mais pas au monument d'intérêt et sont donc mis de côté lors de l'entraînement du modèle SVM.

D'autre part, les points situés sur la frontière visuelle des deux masques de sélection des données peuvent être attribués de façon incertaine à l'une ou l'autre classe. Nous proposons donc d'exclure ces points des données d'entraînement du modèle SVM de façon à ne pas inclure des données incorrectes dans l'une ou l'autre classe. Ces points particuliers sont représentés en bleu sur l'exemple **Figure 39 a)**.

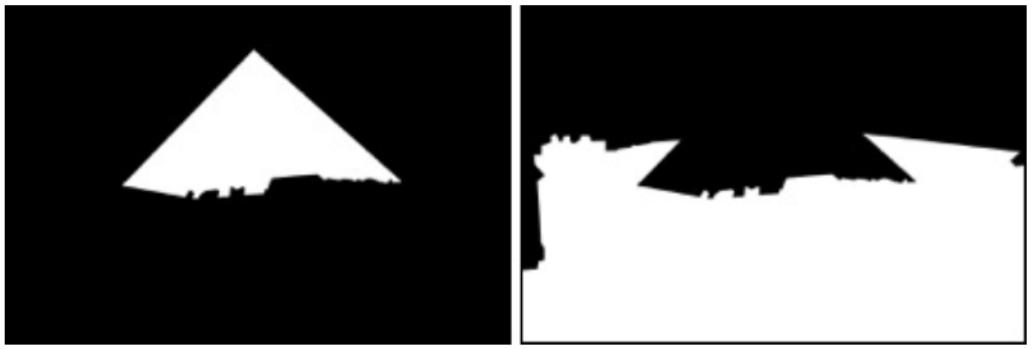
Au final, nous définissons quatre groupes de données pour l'entraînement du classifieur SVM binaire :

- Le premier groupe concerne les descripteurs extraits du monument recherché (représenté par des points verts dans les images prises en exemple **Figure 39** et **Figure 40**).
- Le deuxième groupe inclut les descripteurs extraits des piétons, des trottoirs, des lampadaires, des véhicules, de la végétation…. représentés par des points rouges dans les images illustrées **Figure 39** et **Figure 40**. C'est ce qu'on appelle *l'environnement volatil*, ou *non-persistent*, qui présente *a priori* une variabilité relativement importante.
- Le troisième groupe comporte les descripteurs extraits des bâtiments voisins du monument recherché mais n'en faisant pas parti (représenté par des points jaune dans les images prises en exemple **Figure 39** et **Figure 40**). C'est ce qu'on appelle *l'environnement persistant*, *i.e.*, les éléments de la scène susceptibles d'apparaître de manière récurrente dans les images, selon les différentes prises de vue utilisées lors de l'acquisition.
- Enfin, le quatrième groupe rassemble les points situés à la frontière de séparation des masques de sélection des données des deux premiers groupes (représenté par des points bleus dans les images prises en exemple **Figure 39** et **Figure 40**).

Pour arriver à ce résultat, il est nécessaire de spécifier manuellement deux masques différents, un premier correspondant à la région support du bâtiment et un second aux parties *non-bâtiment* de l'image considérées. Il est à noter que ces deux masques ne sont pas obligatoirement complémentaires comme illustré sur les images **b)** et **c)** des **Figures 39** et **40**. En effet, les bâtiments environnants peuvent inclure des informations utiles pour la définition d'autres bâtiments de la base de données. Ces points d'intérêts ne peuvent donc pas être associés à l'une ou l'autre classe et sont par conséquent exclus de l'entraînement de façon à ne pas biaiser les résultats. En ce qui concerne les masques correspondants aux autres bâtiments qui puissent être présents dans la scène, ils sont par définition le complément dans l'image de la réunion des deux masques *bâtiment* et *non-bâtiments* et illustrés en jaune sur les images a) des **Figures 39** et **40**. D'autre part, les points extraits à la jonction des deux masques initiaux ne sont pas non plus discriminants pour l'une ou l'autre classe. Ces points sont illustrés **Figure 39 a)**. Toujours dans l'optique d'influencer au minimum l'entraînement du modèle SVM, ces points sont aussi exclus des données d'entraînement. Finalement, seuls les points verts et rouges sont conservés pour l'entraînement d'un modèle SVM.



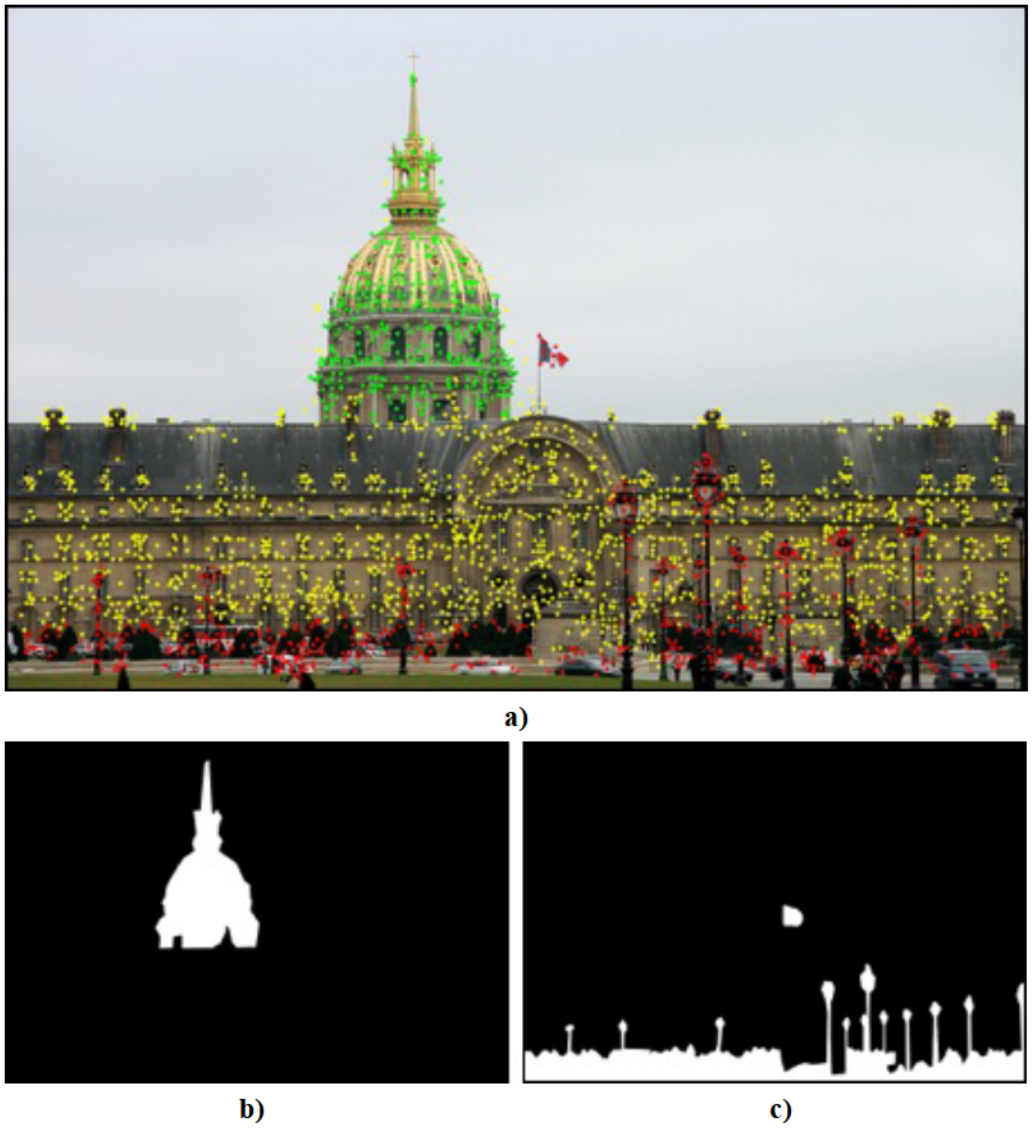
a)



b) **c)**

Figure 39. Exemple de masques de sélection de données pour l'entraînement d'un modèle SVM de classification sur une image du Louvre de la base de données Paris6k. **a)** Visualisation des différentes

classes de points pour l'entraînement, **b)** masque de sélection des points d'entraînement pour la classe *bâtiment*, **c)** masque de sélection des points d'entraînement pour la classe *non-bâtiment*.



a)

b) **c)**

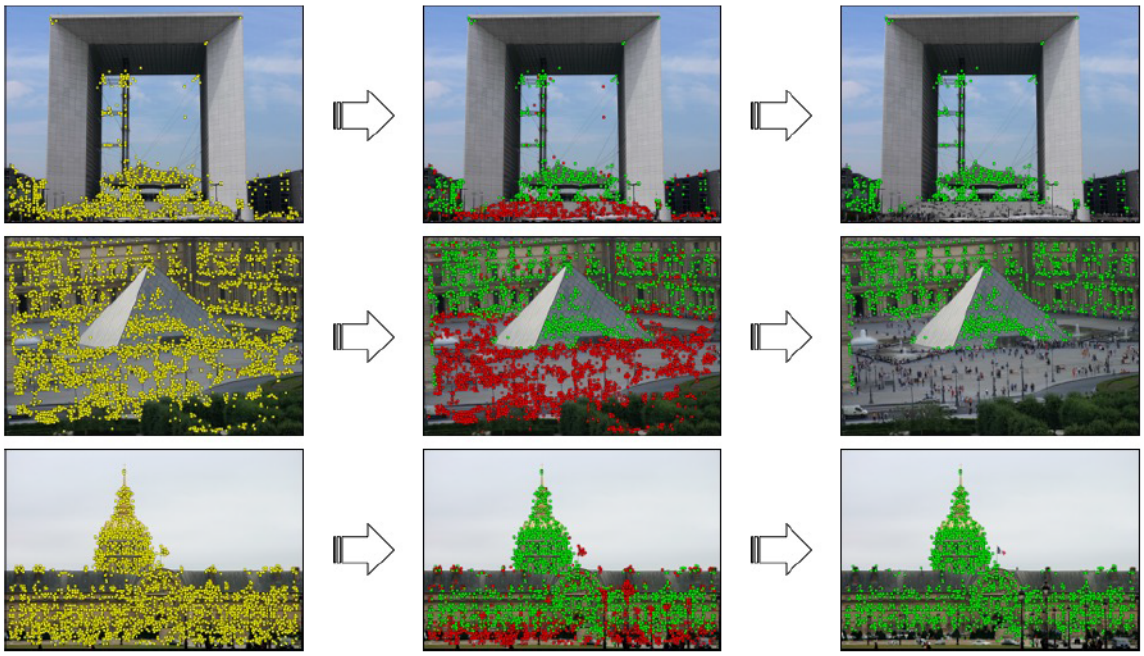
Figure 40. Exemple de masques de sélection de données pour l'entraînement d'un modèle SVM de classification sur une image des Invalides de la base de données Paris6k. **a)** Visualisation des différentes classes de points pour l'entraînement, **b)** masque de sélection des points d'entraînement pour la classe *bâtiment*, **c)** masque de sélection des points d'entraînement pour la classe *non-bâtiment*.

Pour entraîner le classifieur SVM, nous ne retenons que les points des deux premiers groupes, c'est-à-dire les points appartenant de façon certaine à la classe *bâtiment* (en vert) et ceux associés assurément à la classe *non-bâtiment* (en rouge). L'entraînement a été ici réalisé de manière globale, sur l'ensemble des 55 images retenues pour le processus d'apprentissage, tous bâtiments confondus.

Les résultats de prédiction de ce modèle SVM ainsi entraîné sont illustrés par les exemples présentés **Figure 41**. Ici, nous avons utilisé un noyau SVM polynomial avec les paramètres suivants : $C = 10$, $r = 0$, $\delta = 5$ et $\gamma = 1$. Soulignons dès à présent que le modèle polynomial présente de meilleurs

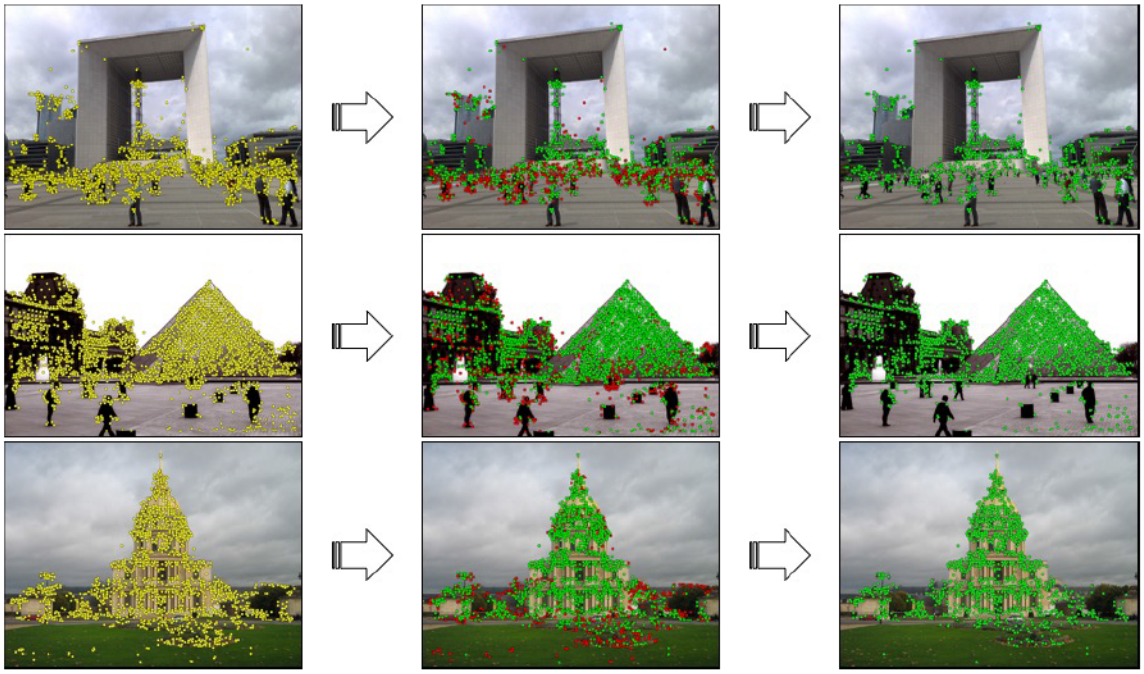
résultats que ceux radial et linéaire. Les résultats expérimentaux obtenus seront présentés en détail et analysés au **Chapitre VII**.

Nous pouvons observer que les points exclus de l'entraînement (notamment des bâtiments environnants, en jaune sur les exemples **Figure 39** et **Figure 40**) sont très majoritairement attribués à la classe *bâtiment* lors de la phase de prédiction. En effet, pour l'exemple de l'image du Louvre illustrée **Figure 39**, les points correspondant aux bâtiments voisins (en jaune) n'ont pas été pris en considération pour la phase d'entraînement. Cependant, comme illustré **Figure 41**, la grande majorité de ces points jaunes sont prédits comme faisant partie de la classe *bâtiment*. Cela peut sembler sans impact particulier sur des images utilisées en entraînement (**Figure 41**), mais pose problème lors de la généralisation de la classification sur des images tests ne faisant pas partie des images d'entraînement du modèle. En effet, comme illustré **Figure 42**, les points détectés au niveau des piétons et d'autres objets de l'environnement sont alors mal catégorisés. La capacité de ce modèle à différencier les descripteurs extraits dans les deux classes *bâtiment* et *non-bâtiment* semble limitée.



a) **b)** **c)**

Figure 41. Exemples de points conservés après filtrage avec un modèle SVM entraînés en excluant les bâtiments voisins des monuments recherchés sur des images utilisées en entraînement : **a)** points originaux extraits, **b)** points répartis dans les classes *bâtiment* (en vert) et *non-bâtiment* (en rouge) et **c)** points *bâtiments* retenus (filtrés).



a) **b)** **c)**

Figure 42. Exemples de points conservés après filtrage avec un modèle SVM entraînés en excluant les bâtiments voisins des monuments recherchés sur des images de test ne faisant pas partie de l'entraînement : **a)** points originaux extraits, **b)** points répartis dans les classes *bâtiment* et *non-bâtiment* et **c)** points *bâtiments* retenus (filtrés).

Dans cette première approche, les points détectés au niveau des bâtiments environnants sont attribués à la classe *bâtiment* lors de l'étape de prédiction. Ce résultat est obtenu malgré la non-inclusion des points détectés au niveau des bâtiments environnants dans les données d'entraînement du modèle de classification. Cela peut s'expliquer par le fait que les différentes catégories de bâtiments sont prises en compte indistinctement lors de l'entraînement du modèle SVM. La prédiction des données issues des bâtiments voisins est alors influencée par les données apportées par d'autres bâtiments de la base de données.

Nous avons alors considéré une deuxième approche, qui consiste à inclure dans le processus d'apprentissage, en tant que points de la catégorie *bâtiment*, les points d'intérêt correspondant à l'environnement persistant.

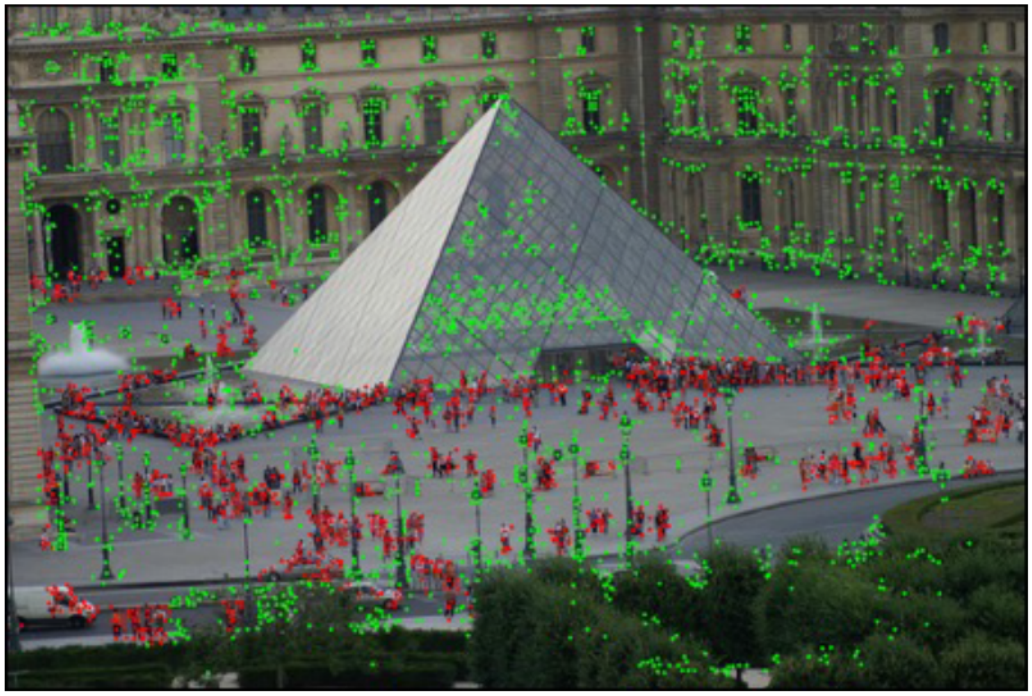
IV.2.1.2. Inclusion de l'environnement persistant

Ces bâtiments, bien que ne faisant pas partie du monument recherché, appartiennent à son environnement et peuvent donc apporter de l'information complémentaire à celle déjà obtenue par le bâtiment d'intérêt seul. De plus, les bâtiments présents dans le voisinage du monument d'intérêt sont visibles dans la majorité des photos du bâtiment recherché et peuvent ainsi aider à identifier le bâtiment recherché. Nous proposons donc d'inclure ces points d'intérêts aux données utiles de la classe *bâtiment* lors de l'entraînement du modèle SVM destiné à faire la distinction entre les

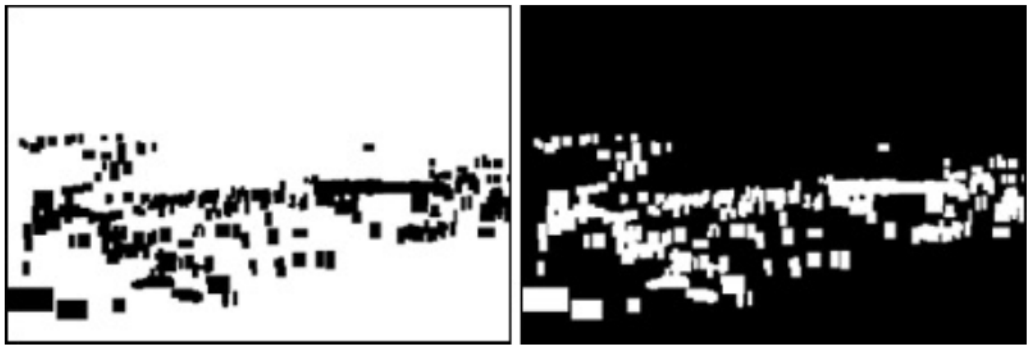
descripteurs pertinents pour la reconnaissance du monument recherché et le bruit provenant des autres objets de la scène.

Des exemples de sélections des données et de leur partitionnement sont illustrés **Figure 43** et **Figure 44**. Tous les descripteurs extraits initialement sont ici pris en considération lors de l'entraînement du modèle SVM, en incluant les points provenant des bâtiments voisins du monument d'intérêt dans la classe *bâtiment*. La classe *non-bâtiment* regroupe ici les objets non persistants de la scène.

Il est à noter que les masques de sélection des données sont dans ce cas complémentaires, contrairement à notre première approche.

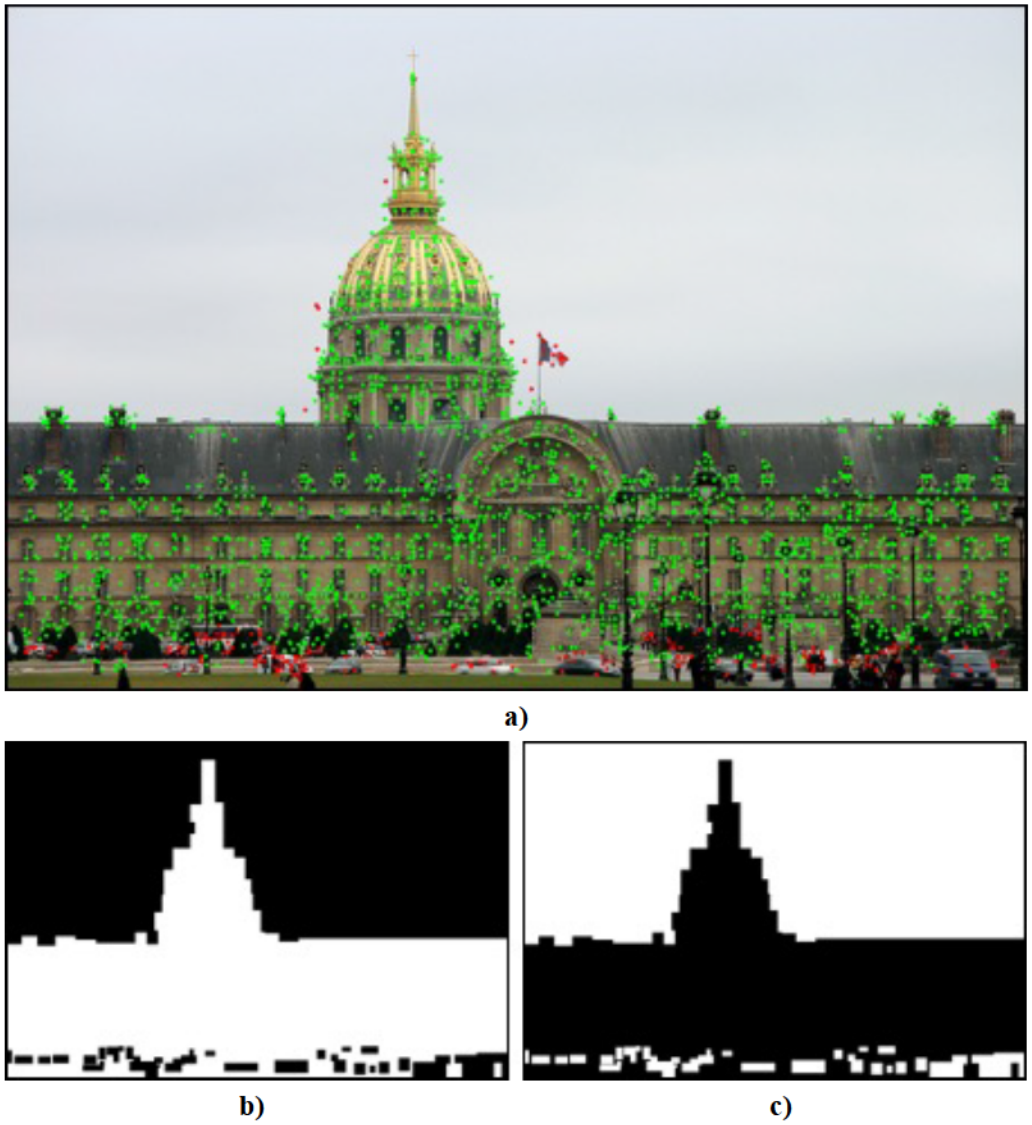


a)



b) **c)**

Figure 43. Exemple de masque de sélection de données pour l'entraînement d'un modèle SVM de classification sur une image du Louvre de la base de données Paris6k. **a)** Visualisation des différentes classes de points pour l'entraînement, **b)** masque de sélection des points d'entraînement pour la classe *bâtiment*, **c)** masque de sélection des points d'entraînement pour la classe *non-bâtiment*.



a)

b) **c)**

Figure 44. Exemple de masque de sélection de données pour l'entraînement d'un modèle SVM de classification sur une image des Invalides de la base de données Paris6k. **a)** Visualisation des différentes classes de points pour l'entraînement, **b)** masque de sélection des points d'entraînement pour la classe *bâtiment*, **c)** masque de sélection des points d'entraînement pour la classe *non-bâtiment*.

Quelques exemples de résultats de prédiction obtenus avec ce modèle SVM sont présentés **Figure 45** (avec des images utilisées lors de l'entraînement) et **Figure 46** (avec des images tests n'ayant pas été prise en compte lors de l'entraînement).

Nous pouvons remarquer dans un premier temps que, dans tous les cas, très peu de points sont attribués à la classe *non-bâtiment*. La quasi-totalité des points initiaux est donc conservée malgré cette étape de filtrage ajoutée. Quelques descripteurs au niveau des piétons sont effectivement rejetés mais

nous constatons que la classe *bâtiment* est largement dominante par rapport à la classe *non-bâtiment*. Cette observation est d'autant plus flagrante sur les images tests non utilisées lors de la phase d'entraînement comme illustré **Figure 46**. De nombreux faux positifs sont notamment visibles au niveau des piétons bien que les points associés soient bien entraînés comme faisant partie de la classe *non-bâtiment*. La sélection des données d'entraînement de la classe *bâtiment* est donc trop vaste. Par conséquent, les points sont attribués trop couramment à cette même classe lors de la prédiction basée sur le modèle de classification SVM.

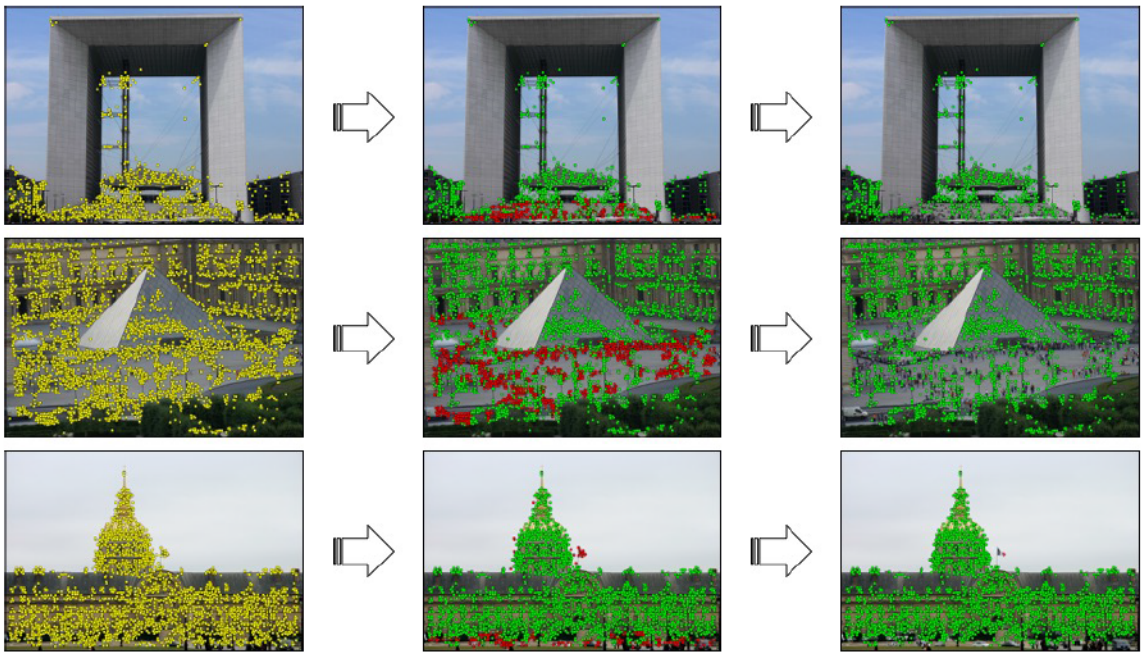
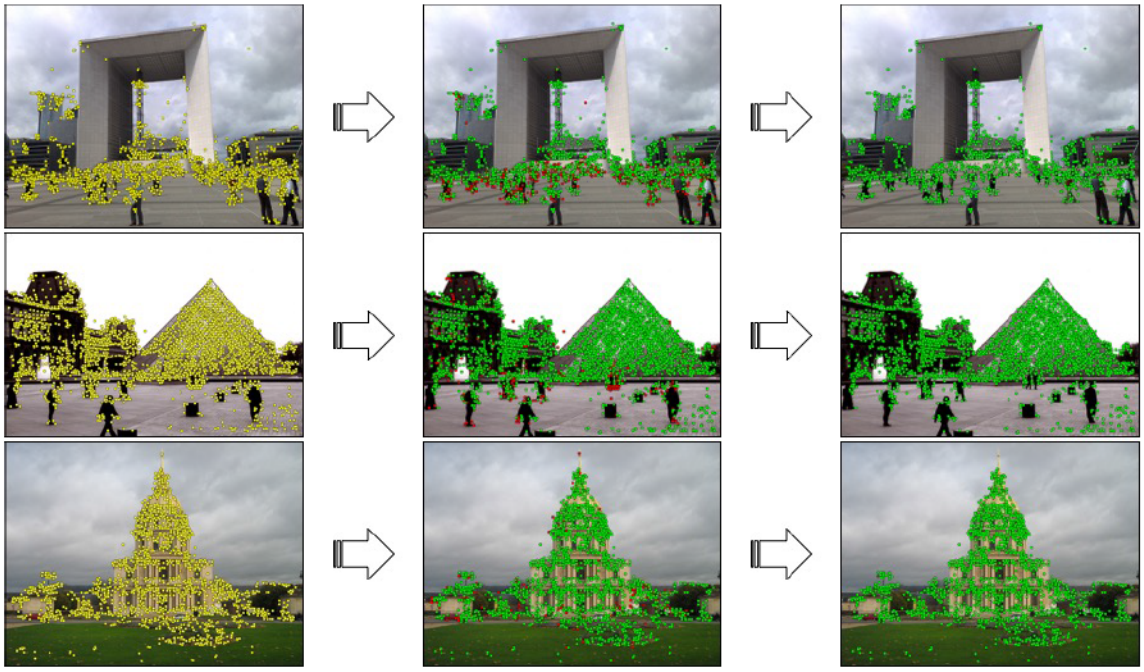


Figure 45. Exemples de points conservés après filtrage avec un modèle SVM entraînés en excluant les bâtiments voisins des monuments recherchés sur des images utilisées en entraînement : **a)** points originaux extraits, **b)** répartis dans les classes *bâtiment* et *non-bâtiment* et **c)** filtrés.



a) b) c)

Figure 46. Exemples de points conservés après filtrage avec un modèle SVM entraînés en excluant les bâtiments voisins des monuments recherchés sur des images de test ne faisant pas partie de l'entraînement : **a)** points originaux extraits, **b)** répartis dans les classes bâtiment et non-bâtiment et **c)** filtrés.

Cette approche ne peut donc pas non plus représenter une solution efficace à notre problème. Afin de s'affranchir des inconvénients susmentionnés, nous avons considéré et testé une troisième stratégie, plus restrictive, qui s'appuie sur une isolation stricte de l'objet d'intérêt. Elle est détaillée dans la section suivante.

IV.2.1.3. Isolation stricte de l'objet d'intérêt

Cette fois, nous proposons de prendre en compte lors de l'entraînement uniquement les points correspondant au bâtiment d'intérêt dans la classe *bâtiment* et d'attribuer à la classe *non-bâtiment* l'ensemble du reste des points d'intérêts détectés. Notamment les descripteurs extraits au niveau des bâtiments voisins du monument recherché sont maintenant considérés comme n'étant pas pertinents pour la reconnaissance *a posteriori* du bâtiment.

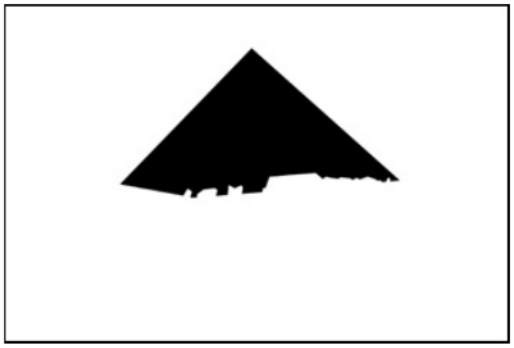
Ce principe est illustré pour les exemples présentés **Figure 47** et **Figure 48**, avec les masques de sélection associés. Il est à noter que ces masques de sélection sont à nouveau complémentaires pour l'une et l'autre des deux classes lors de l'entraînement du modèle SVM. En effet, nous avions remarqué précédemment que même en excluant les bâtiments environnants de l'entraînement, les points détectés à leur niveau sont tout de même associés à la classe *bâtiment* lors de la phase de prédiction.

Notre objectif ici est de limiter notre filtrage aux données issues strictement du monument d'intérêt de façon à ne conserver que les descripteurs les plus pertinents et représentatifs dudit bâtiment. Bien que certains objets fassent partie de l'environnement persistant du monument,

recherché, comme nous l'avons vu dans le paragraphe précédent, ces bâtiments peuvent être confondus avec d'autres monuments d'intérêts de la base de données. Cela ajoute donc une certaine confusion et plutôt que d'apporter de l'information utile, amplifie en réalité le "bruit" dans la classification des descripteurs.

a)

b)

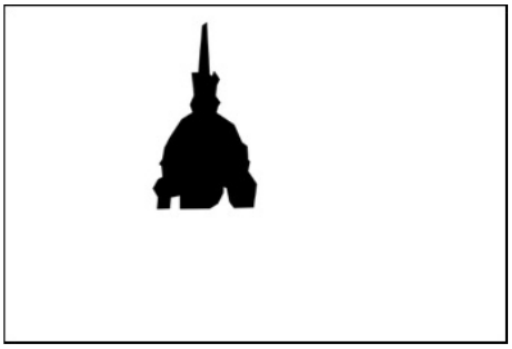


c)

Figure 47. Exemple de masque de sélection de données pour l'entraînement d'un modèle SVM de classification sur une image du Louvre de la base de données Paris6k. **a)** Visualisation des différentes classes de points pour l'entraînement, **b)** masque de sélection des points d'entraînement pour la classe bâtiment, **c)** masque de sélection des points d'entraînement pour la classe non-bâtiment.

a)

b)



c)

Figure 48. Exemple de masque de sélection de données pour l'entraînement d'un modèle SVM de classification sur une image des Invalides de la base de données Paris6k. **a)** Visualisation des différentes classes de points pour l'entraînement, **b)** masque de sélection des points d'entraînement pour la classe bâtiment, **c)** masque de sélection des points d'entraînement pour la classe non-bâtiment.

Quelques résultats obtenus avec ce modèle SVM de classification sont présentés **Figure 49** (pour des images utilisées en entraînement) et **Figure 50** (pour des images tests n'ayant pas été prises en compte lors de cet entraînement).

Nous remarquons dans un premier temps que les images utilisées en entraînement filtrent de façon quasi-parfaite les descripteurs extraits du bâtiment d'intérêt. En effet, les descripteurs conservés après filtrage correspondent quasiment en totalité aux points définis dans la classe *bâtiment* lors de l'entraînement du modèle.

Cependant, la généralisation lors des tests effectués avec des images non prises en compte lors du processus d'apprentissage montre des erreurs de prédiction et notamment toujours un nombre non négligeable de faux positifs au niveau de la végétation et des piétons, comme illustré **Figure 50**. Nous illustrons les résultats obtenus avec ce modèle avec d'autres images de la base de données Paris6k dans la **Section IV.4**.

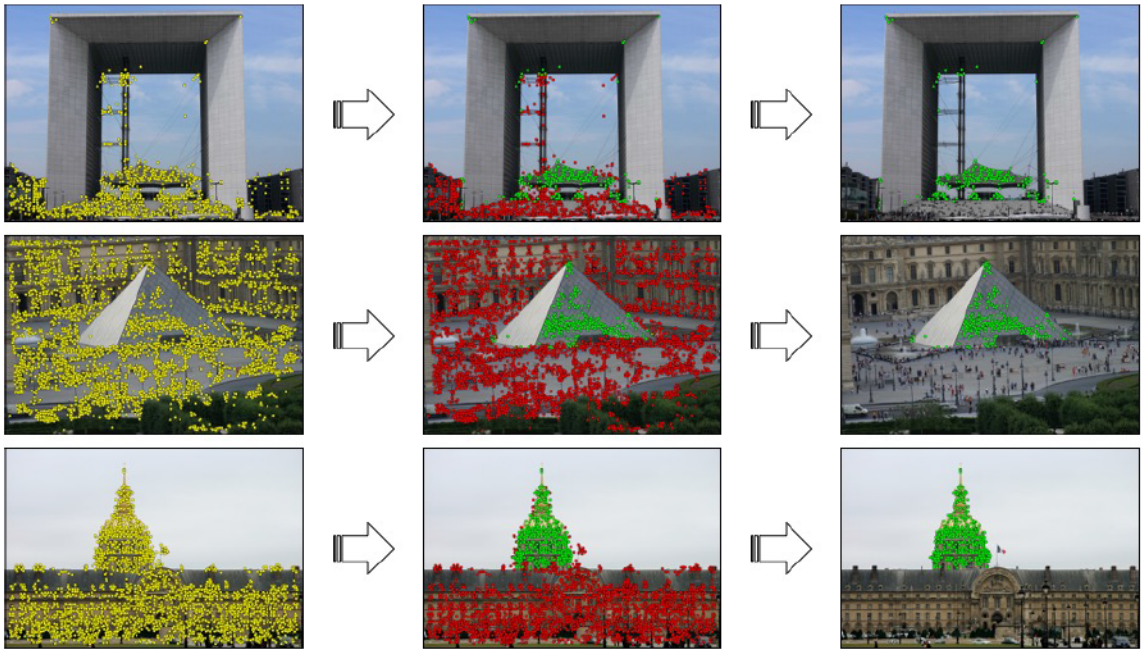


Figure 49. Exemples de points conservés après filtrage avec un modèle SVM entraînés en excluant les bâtiments voisins des monuments recherchés sur des images utilisées en entraînement pour une image dans la base d'entraînement : **a)** points originaux extraits, **b)** répartis dans les classes *bâtiment* et *non-bâtiment* et **c)** filtrés.

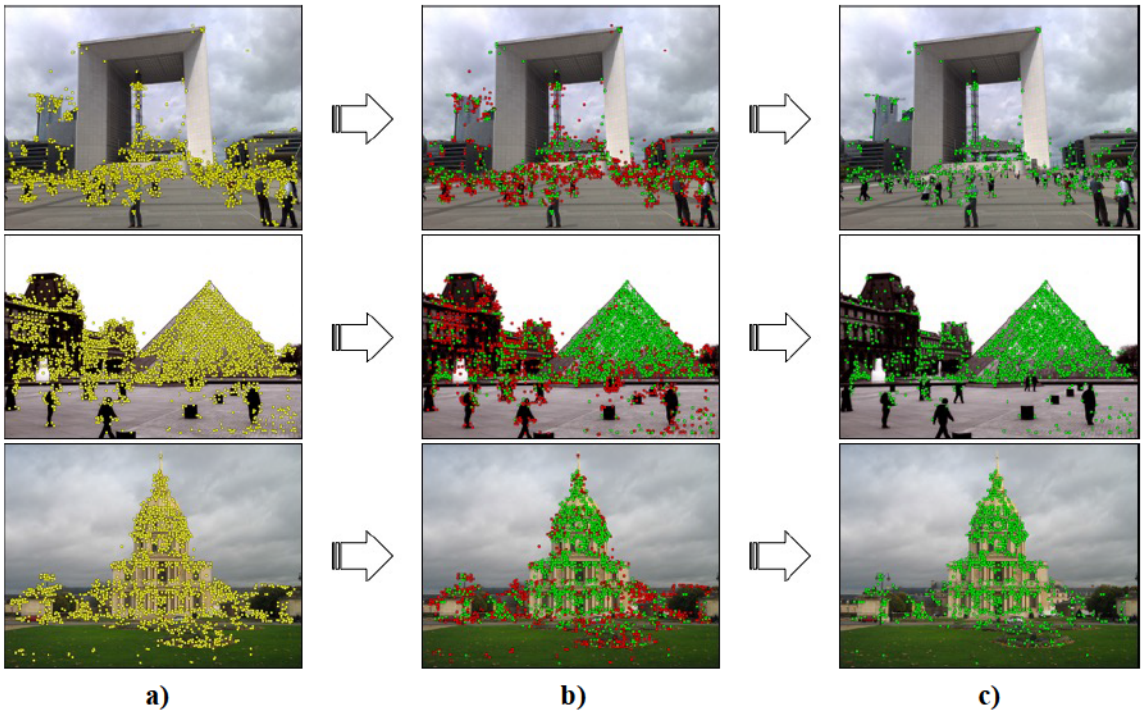


Figure 50. Exemples de points conservés après filtrage avec un modèle SVM entraînés en excluant les bâtiments voisins des monuments recherchés sur des images de test ne faisant pas partie de l'entraînement : **a)** points originaux extraits, **b)** répartis dans les classes *bâtiment* et *non-bâtiment* et **c)** filtrés.

IV.3. Paramètres d'apprentissage

Les classifieurs SVM sont entraînés sur la base des images définies en tant qu'images requêtes dans la vérité terrain. Dans les deux bases de données Paris6k et Oxford5k considérées ici, la vérité terrain compte 11 catégories de bâtiments et 5 images pour chacune des catégories sont retenues en tant qu'images requêtes.

L'utilisation d'une technique de classification SVM nécessite de choisir un ensemble de paramètres adaptés à notre problématique. Dans notre cas, ce choix a été guidé par une étude empirique visant à établir l'influence de chaque paramètre sur les performances de la classification. Ces éléments d'analyse sont présentés dans les paragraphes suivants.

IV.3.1. Type de noyau du modèle SVM

Lorsqu'on considère un processus d'apprentissage par modèle SVM, le premier élément à retenir concerne la spécification de noyau du modèle. Différents choix s'offrent à nous, incluant de modèles linéaire, radial ou encore polynomiaux de différents degrés (**Figure 51**).

Dans un premier temps, nous avons considéré comme cas d'école un exemple simple, de classification d'un ensemble de points dans le plan 2D. Ces expériences sont réalisées sur une machine de 16GB de RAM, 8 cœurs sous système d'exploitation Ubuntu version 14.04. Les résultats de classification obtenus différents types de noyaux sont illustrés. Pour le noyau polynomial, nous avons considéré un degré 5.

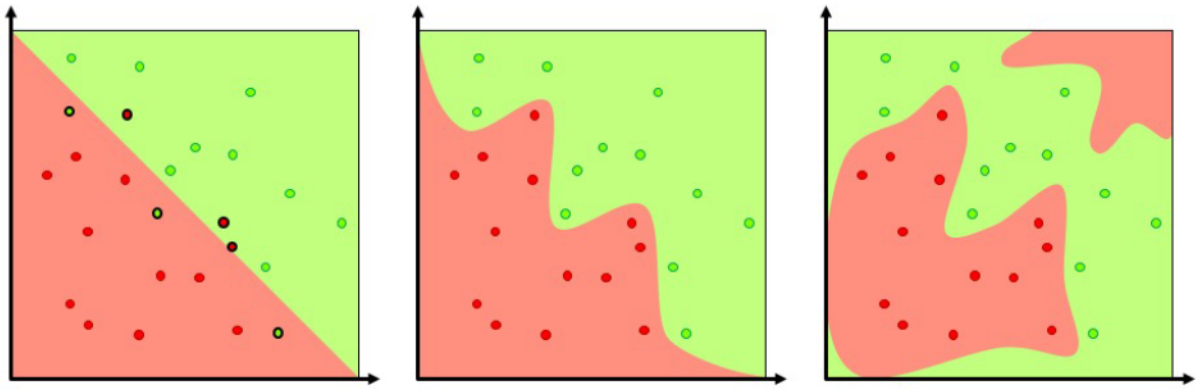


Figure 51. Exemples d'entraînement SVM avec différents types de noyau **a)** linéaire **b)** polynômial de degré 5 **c)** radial.

A partir de ce cas d'école, nous pouvons déjà identifier la difficulté du modèle linéaire de séparer les deux classes, de faux positifs et de faux négatifs apparaissant à la fois. Les modèles polynomial et radial arrivent à s'affranchir de cette limitation et à séparer correctement les points des

deux classes, avec un léger avantage pour le modèle polynomial qui évite de partager l'espace en multiples composantes connexes.

La **Figure 52** présente les résultats de classification obtenus sur les images de la vérité terrain proposées dans la base de données Paris6k.

Le modèle SVM à noyau linéaire prédit la quasi-totalité des données dans une unique classe. Les types de noyaux polynomial et radial montrent des avantages différents. Le taux de faux positifs est réduit dans un noyau de type radial mais le pourcentage de données attribuées correctement à leur classe est plus favorable en moyenne avec un noyau polynomial.

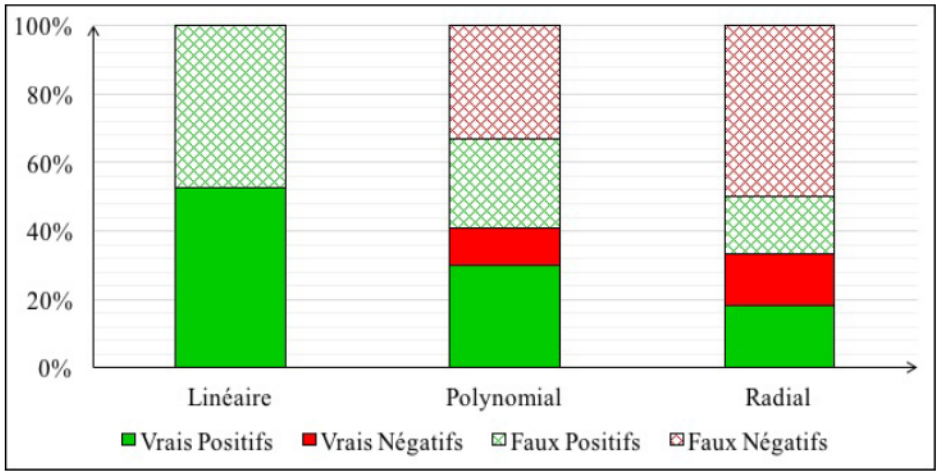


Figure 52. Résultats de la classification SVM en fonction du type de noyau choisi.

Pour ces raisons, nous avons retenu un modèle de classification de type polynomial. Quant au degré du polynôme retenu, son choix est explicité dans le paragraphe suivant.

IV.3.2. Degré du polynôme définissant l'hyperplan de séparation et terme d'interception r

Dans le cas spécifique d'un noyau du modèle SVM polynomial, deux autres paramètres sont à prendre en compte. Il s'agir notamment du degré du polynôme considéré et du terme d'interception *r* (*cf.* équation (23), **Section IV.1**).

Le degré du polynôme définissant l'hyperplan de séparation des données permet de définir la complexité de cette frontière. Intuitivement, plus le degré du polynôme est élevé, plus la frontière de séparation peut être complexe et moins il y a de risque d'erreur de classification comme illustré **Figure 53** ainsi que dans les exemples présentés **Figure 54**.

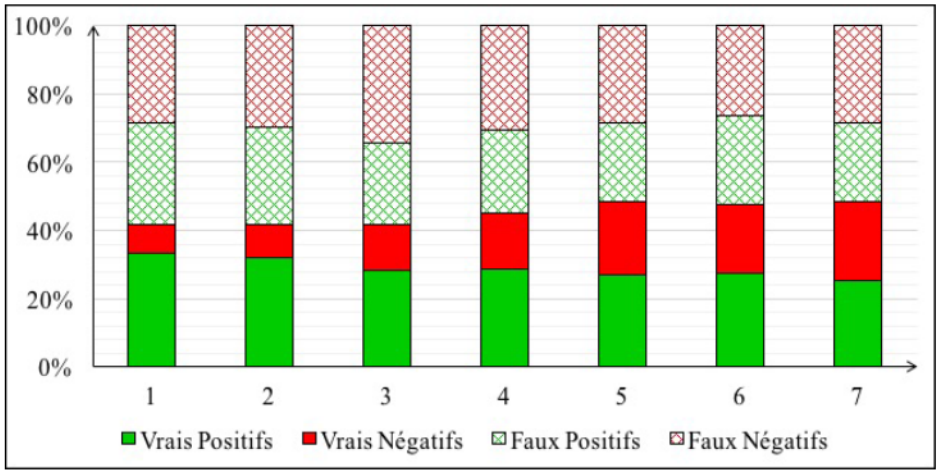


Figure 53. Résultats de la classification SVM en fonction du degré du polynôme δ dans le cas d'un noyau polynomial.

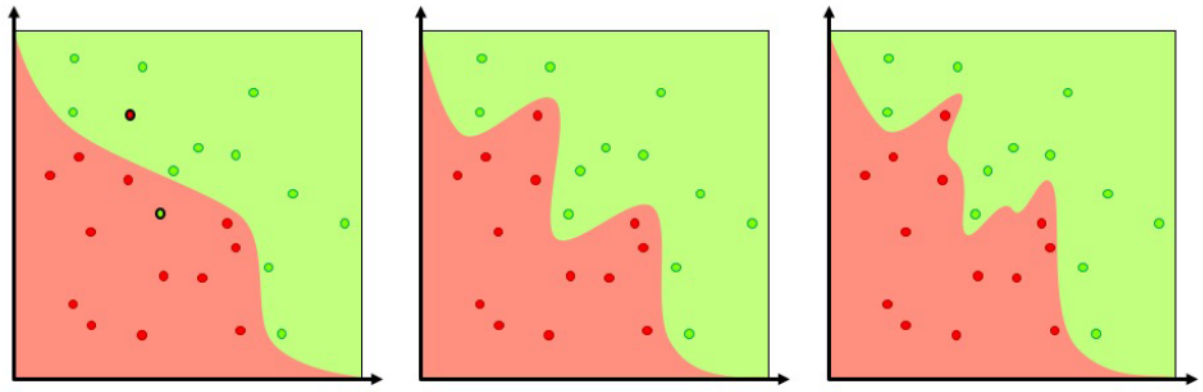


Figure 54. Exemples d'entraînements SVM avec différents polynômes de degrés différents **a)** δ = 3 **b)** $\delta = 5$ **c)** $\delta = 7$.

La **Figure 53** présente les résultats obtenus par classification sur les images de la vérité terrain de la base Paris6k. Nous pouvons observer que le taux de données correctement classifiées en prédiction avec un modèle d'entraînement basé sur un noyau de type polynômial augmente avec le degré du polynôme. Cependant, une frontière plus complexe nécessite un temps de calcul plus important sans pour autant améliorer considérablement la précision de séparation des deux espaces de classifications.

En ce qui concerne notre cas d'école avec ensembles de points 2D, les résultats sont illustrés **Figure 54**, pour de noyaux polynomiaux de degrés 3, 5 et 7. Un faux positif est obtenu uniquement pour le polynôme cubique (**Figure 54 a)**). Pour les degrés 5 et 7 (**Figure 54 b)** et **Figure 54 c)**), la classification est correcte dans les deux cas, avec une frontière relativement plus complexe dans le cas du degré 7.

Nous retenons pour nos développements un degré de polynôme $\delta = 5$, qui présente un bon équilibre entre temps de calcul et résultats de classification.

En ce qui concerne le terme d'interception r, il n'influe pas sur les performances de précisions des différents modèles SVM, bien que nous ayons testé des valeurs dans un spectre relativement large, comme illustré **Figure 55**.

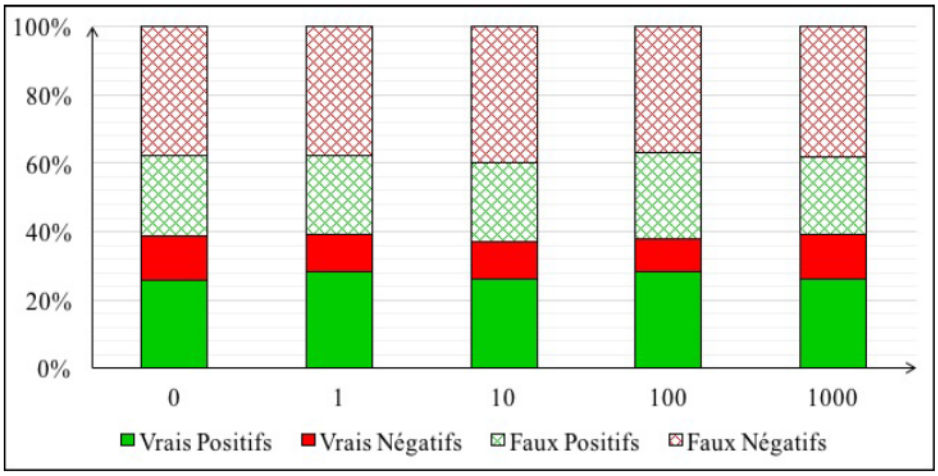


Figure 55. Résultats de la classification SVM en fonction du terme d'interception r dans le cas d'un noyau polynomial.

Nous avons donc gardé la valeur par défaut de $r = 0$.

IV.3.3. Paramètre de pénalisation C

Le paramètre de pénalisation C (équation (22), **Section IV.1**) contrôle une tolérance plus ou moins importante aux erreurs de classification lors de la phase d'entraînement.

La **Figure 56** illustre l'influence de ce paramètre pour des valeurs échelonnées entre 0,01 et 100. Un maximum de données correctement attribuées est obtenu pour une valeur de $C = 0{,}01$. Cependant, dans ce cas la majorité des données sont prédites dans une unique classe et nous pouvons observer un grand taux de faux positifs. Le meilleur équilibre est obtenu pour une valeur de $C = 10$. Pour des valeurs supérieures à 10, les résultats se dégradent.

Pour les grandes valeurs de C, l'optimisation construit un hyperplan de séparation avec une marge plus étroite dans le but d'obtenir tous les points d'entraînement classés correctement comme illustré **Figure 57 a)**. A l'inverse, une très faible valeur du paramètre C permet au modèle de construire un hyperplan de séparation simplifié, même si pour cela des données sont volontairement mal classées comme illustré **Figure 57 b)**. Dans cet exemple, avec une valeur du paramètre C élevée, toutes les données d'entraînement sont correctement classifiés alors qu'avec une valeur faible, les points entourés de noir sont volontairement mal classifiés pour l'entraînement dans le but de simplifier la définition de l'hyperplan de séparation de l'espace.

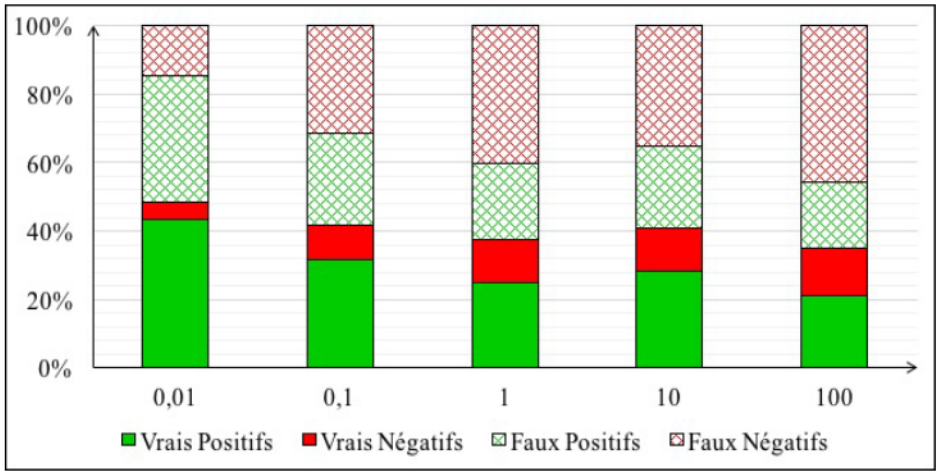


Figure 56. Résultats de la classification SVM en fonction du paramètre de pénalisation *C*.

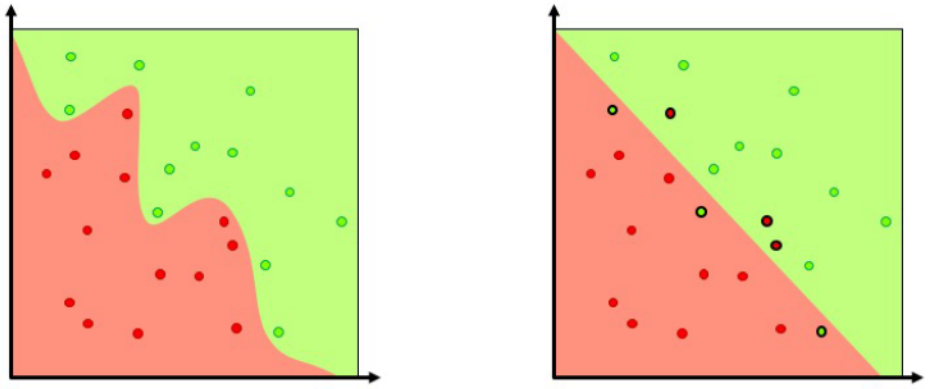


Figure 57. Exemples d'entraînement SVM avec différents paramètres C **a)** avec une valeur élevée **b)** avec une valeur faible.

IV.3.4. Paramètre d'influence γ

Un des paramètres influant grandement l'entraînement d'un modèle SVM est le paramètre γ. L'influence de ce paramètre est illustré **Figure 58** pour des valeurs de γ variant de 1×10^{-12} à 1 000. Pour des valeurs extrêmes, la quasi-totalité des données sont classifiées dans une unique classe. L'influence des données d'une des deux classes *bâtiment* et *non-bâtiment* est donc prépondérante par rapport à l'autre.

Une faible valeur de γ implique que la région d'influence des vecteurs supports s'étend à la quasi-totalité des données d'entraînement comme illustré **Figure 59 a)**. Dans ce cas, l'influence de la classe *verte* sur l'entraînement prend l'ascendant sur l'autre classe.

Dans le cas contraire d'une valeur de γ élevée illustrée **Figure 59 b)**, l'influence des vecteurs de support de la classe *verte* se limite aux données elles-mêmes. Dans l'exemple donné ici, aucune donnée n'est mal classifiée en entraînement, mais les points utilisés en entraînement de la classe *verte* sont les seuls points définissant l'espace de cette classe pour ce modèle de classification. Bien

qu'aucune donnée ne soit mal classifiée en entraînement, lors de la phase de prédiction de données inconnues, aucune marge d'erreur n'est possible. Il s'agit donc de sélectionner une valeur permettant un équilibre. Dans notre cas, en prenant en compte les résultats présentés **Figure 58**, nous avons opté pour une valeur $\gamma = 10$.

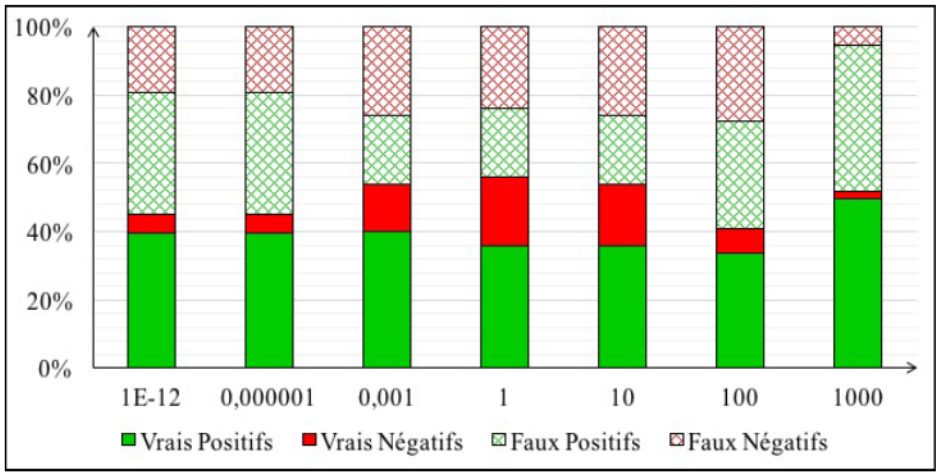


Figure 58. Résultats de la classification SVM en fonction du paramètre γ.

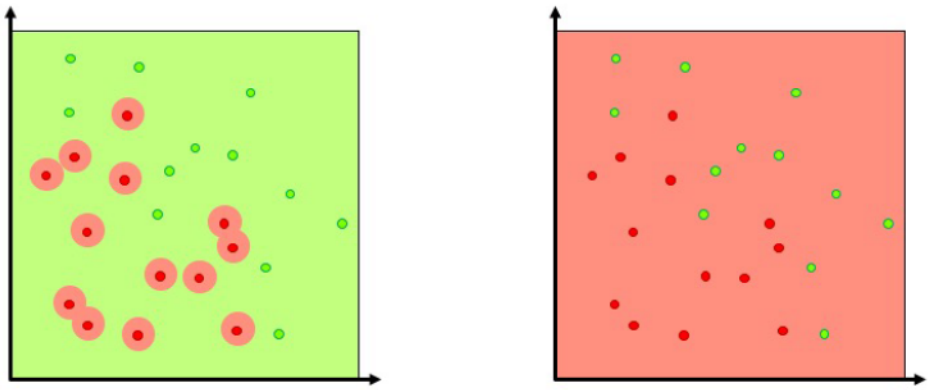


Figure 59. Exemples d'entraînement SVM avec différents paramètres γ **a)** avec une valeur faible **b)** avec une valeur élevée.

Cela conclut notre analyse sur le choix des paramètres des modèles SVM. En résumé, nous avons retenu un noyau SVM de type polynomial avec le paramètre de pénalisation $C = 10$, le terme d'interception $r = 0$, le degré du modèle polynomial $\delta = 5$ et enfin un paramètre d'influence $\gamma = 1$. Ces paramètres seront conservés pour l'ensemble des résultats présentés dans cette thèse.

Présentons maintenant et analysons les résultats obtenus sur différentes images de la base Paris6k, à la fois d'entraînement et de test.

IV.4. Résultats de la classification sur les images d'entraînement et test

Nous présentons dans cette section les résultats visuels de la classification SVM avec les paramètres définis précédemment. Pour les catégories présentées ici, les résultats sont donnés à chaque fois pour une image utilisée lors de la phase d'entraînement et pour une image utilisée uniquement lors de la phase de test. Dans les **Figures 60**, **61** et **62** les points verts correspondent aux points attribués à la catégorie *bâtiment* et les points rouges à la catégorie *non-bâtiment*.

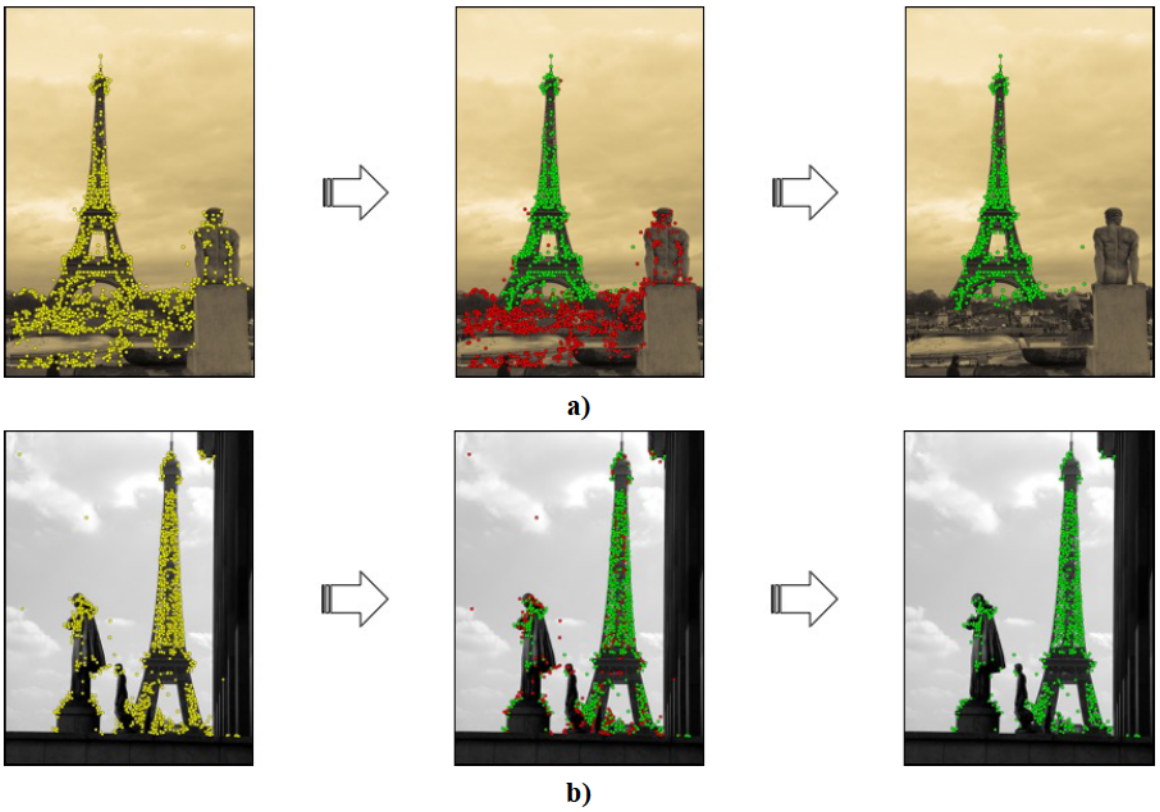


Figure 60. Résultats de la classification de descripteurs SIFT pour la catégorie Tour Eiffel **a)** pour une image d'entraînement **b)** pour une image test.

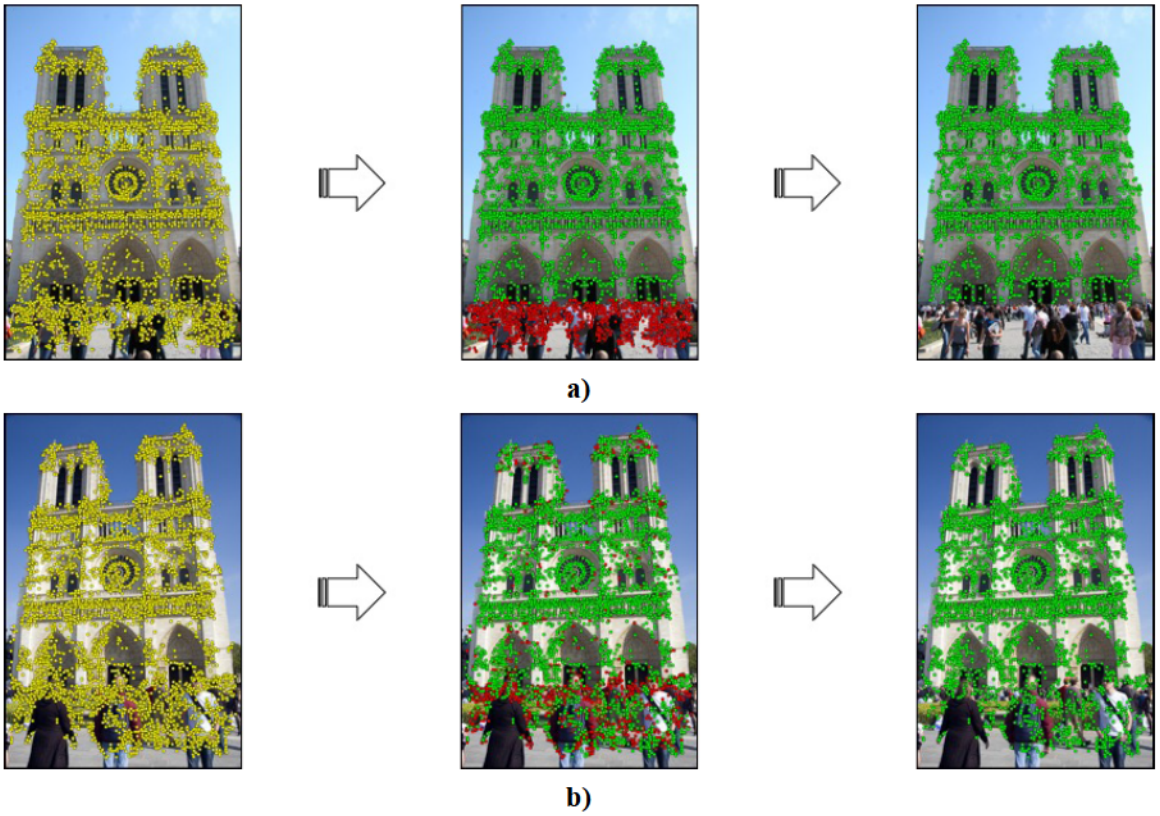


Figure 61. Résultats de la classification de descripteurs SIFT pour la catégorie Notre Dame **a)** pour une image d'entraînement **b)** pour une image test.

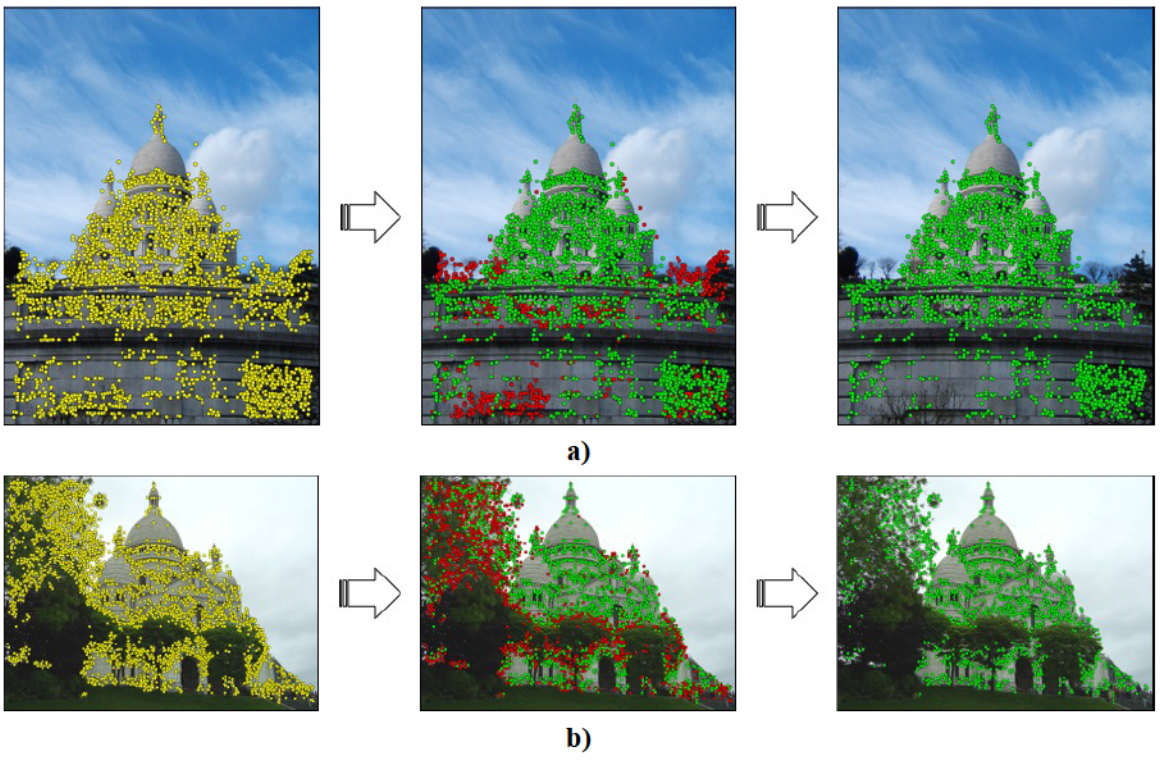


Figure 62. Résultats de la classification de descripteurs SIFT pour la catégorie Sacré Cœur **a)** pour une image d'entraînement **b)** pour une image test.

Dans les **Figures 60** et **61**, les images utilisées en entraînement montrent bien une classification quasi parfaite des descripteurs. Cependant, la généralisation aux images de test présente des résultats plus disparates et notamment de nombreux faux positifs.

Dans le cas de l'exemple de la **Figure 62**, l'image utilisée en entraînement montre déjà des résultats hasardeux avec ici encore de nombreux faux positifs. De par le fait que nous entraînons un seul classifieur SVM global pour toutes les catégories ce classifieur n'est pas complètement adapté à chaque bâtiment. La végétation et les piétons sont bien attribués à la classe *non-bâtiment* mais la classification est moins fiable pour les bâtiments environnants. Ainsi, même pour une image utilisée en entraînement, les descripteurs extraits d'un mur devant le bâtiment recherché sont attribués à la classe *bâtiment* comme illustré sur la **Figure 62 a**).

La classification erronée de ces points est à notre avis due à la façon d'entraîner le modèle SVM sans distinction des différentes catégories de bâtiments de la base de données. Néanmoins, cette dernière approche est de loin la plus sélective et, dans le cadre d'une stratégie d'apprentissage globale, aboutit au filtrage le plus fin possible des points issus du bâtiment recherché.

Cette analyse nous conduit à considérer une autre stratégie, qui abandonne le paradigme d'un apprentissage global. Dans les bases de données utilisées pour la recherche d'images, les images des bâtiments sont regroupées dans différentes catégories précisées dans la vérité terrain. Ainsi, il est possible d'entraîner un modèle SVM différent pour chaque catégorie spécifiée dans la vérité terrain.

Dans ce cas, la classe *bâtiment* est spécifique à chaque catégorie et le classifieur est donc adapté à chaque bâtiment. Cependant nous devons donc choisir le classifieur correct pour chaque image sans aucune connaissance préalable de la catégorie à laquelle elle appartient. Cela signifie qu'une étape de fusion est nécessaire afin de déterminer le classifieur le plus adapté.

Cette approche d'apprentissage par modèles SVM adaptés aux catégories de bâtiments est détaillée dans le chapitre suivant.

V. MODELES SVM ADAPTES PAR CLASSE DE BATIMENTS

Résumé. Ce chapitre introduit un des éléments centraux de notre travail, qui concerne l'élaboration d'un modèle de classification multiple, pouvant prendre en compte l'ensemble des catégories à détecter pour une base de données avec vérité terrain. Cela implique la construction d'un classifieur indépendant pour chacune des catégories considérées. Dans un premier temps, nous nous concentrons sur les modalités permettant d'affiner la classification binaire des points d'intérêt de type *bâtiment* ou *non bâtiment*. Ensuite, nous nous consacrons à la spécification d'une stratégie de classification globale, pouvant identifier, pour une image donnée, le classifieur le plus adapté. Cette deuxième étape implique un processus de fusion *a posteriori* des classifieurs.

Dans ce cadre, deux critères différents sont proposés et évalués. Un premier s'appuie simplement sur le nombre de points classifiés dans une catégorie donnée. Le second critère affine cette première stratégie en prenant en compte une mesure de vraisemblance globalisée.

Les résultats expérimentaux obtenus sont présentés à la fois pour la classification des descripteurs locaux que pour les stratégies globales de sélection d'un classifieur adapté à chaque image. Pour les deux bases de données retenues pour évaluation ils démontrent la supériorité de l'approche par classifieurs multiples, en comparaison avec l'approche de classification globale présentée au **Chapitre IV**.

Mots clés : classification de données, modèles SVM multi-classes, modèles adaptés.

Nous introduisons dans ce chapitre un modèle de classification multiple des descripteurs locaux qui s'appuie sur une série d'entraînements SVM adaptés à chaque catégorie de la base de données.

V.1. Apprentissage multi-classe adapté

Différentes catégories de bâtiments étant définies dans la vérité terrain des bases de données, il est donc possible d'entraîner un modèle SVM différent pour chaque catégorie. De cette manière, chaque modèle est adapté au bâtiment recherché dans sa catégorie. La **Figure 63** illustre le processus d'entraînement d'un unique classifieur global sans distinction des différentes catégories définies dans la base de données.

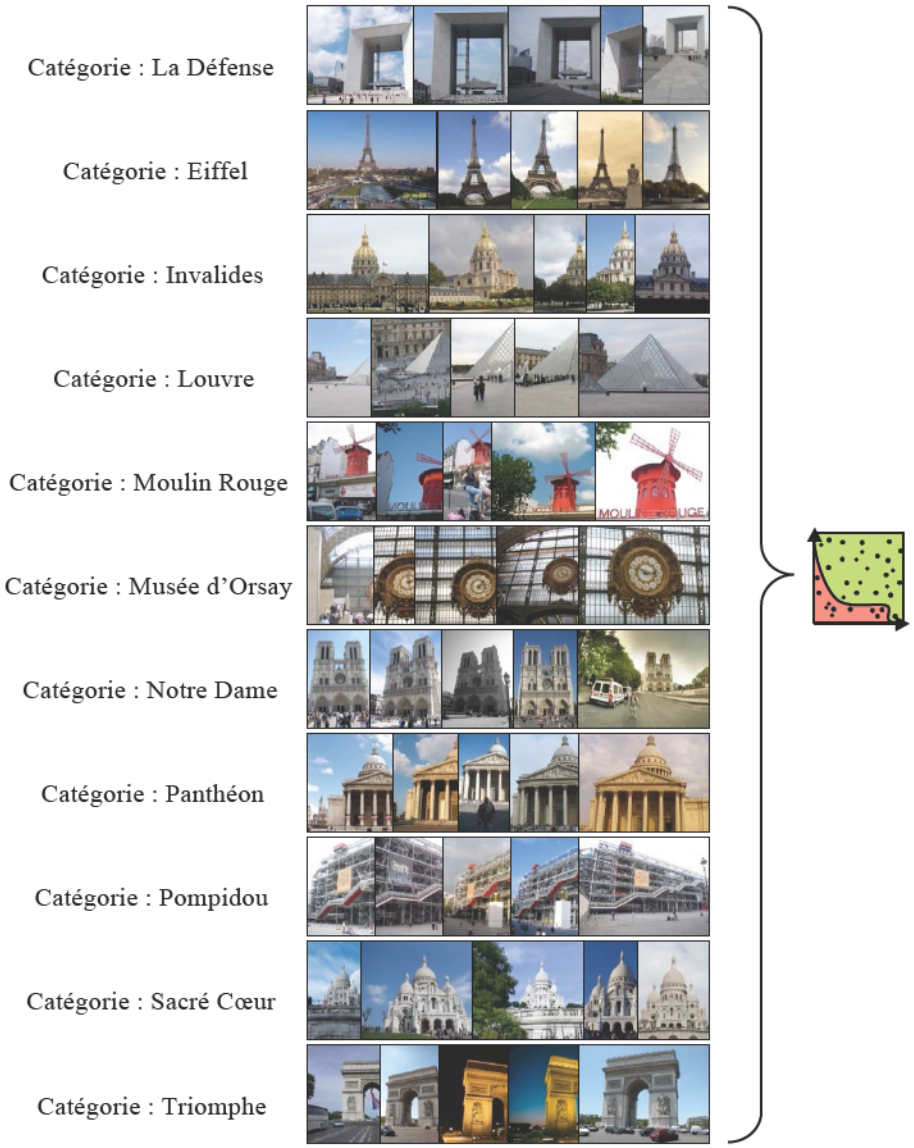


Figure 63. Illustration d'un unique classifieur SVM global entraîné indistinctement des différentes catégories de bâtiments de la base de données Paris6k.

Nous proposons à présent un entraînement spécifiquement adapté à chaque catégorie définie dans la vérité terrain de la base de données. Cette nouvelle stratégie est illustrée **Figure 64**, où 11 modèles SVM dédiés sont entraînés spécifiquement pour chacun des monuments de la base de données Paris6k.

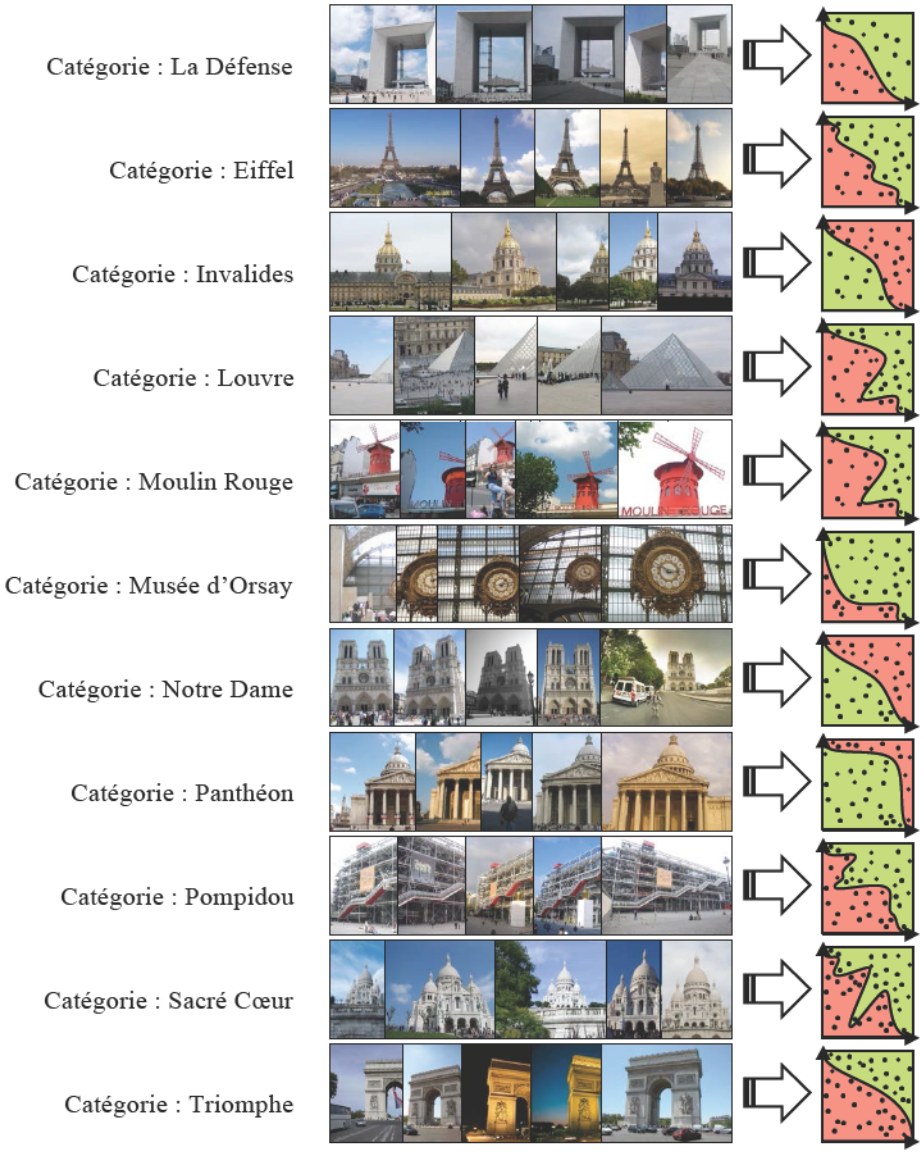


Figure 64. Illustration des différents classifieurs adaptés à chaque catégorie de bâtiment de la base de données Paris6k.

Pour mettre en œuvre une telle approche, il est nécessaire de définir de manière rigoureuse les données utilisés pour l'entraînement avec à la fois de exemples positifs et négatifs. Plusieurs choix sont alors possibles. La section suivante présente notamment les différentes possibilités de sélection des données d'entraînement.

V.1.1. Définition des données d'entraînement

Nous avons envisagé et testé plusieurs approches lors de la sélection des données utilisées en entraînement.

V.1.1.1. Entraînement indépendant par catégorie de bâtiment

En s'appuyant sur les résultats obtenus et rapportés au chapitre précédent dans le cas de l'approche de classification globale (**Section IV**), nous proposons maintenant d'entraîner un classifieur spécifique pour chaque catégorie de bâtiment comme illustré **Figure 64**.

Par exemple, pour la base de données Paris6k, 11 catégories de bâtiments sont identifiées (la Défense, la Tour Eiffel, les Invalides, le Louvre, le Moulin Rouge, le Musée d'Orsay, Notre Dame, le Panthéon, le musée Pompidou, le Sacré Cœur et l'arc de Triomphe) et 5 images sont proposées pour entraîner les 11 modèles SVM associés.

Pour ce faire, la sélection des données d'entraînement est réalisée comme décrit précédemment (**Section IV**) en créant des masques de sélection pour chacune des 5 images d'entraînement. La principale différence est que dans ce cas 11 modèles SVM adaptés sont générés au lieu d'un unique modèle global.

V.1.1.2. Entraînement discriminant d'une catégorie de bâtiment par rapport aux autres

Les images utilisées en entraînement sont limitées à 5 par catégorie pour les données représentant la classe *bâtiment* d'une catégorie donnée. Cependant, de façon à ce qu'un bâtiment soit distingué de façon efficace d'un autre bâtiment, les 50 images des 10 bâtiments restants sont attribuées à la classe *non-bâtiment* de la catégorie donnée.

Par exemple, dans la catégorie Tour Eiffel, les descripteurs issus des bâtiments de la Défense sont attribués à la catégorie *non-bâtiment* de façon à distinguer le plus précisément possible la Tour Eiffel de l'arche de La Défense. Dans cet exemple, nous illustrons les données considérées comme faisant partie de la classe *bâtiment* pour l'entraînement sur la **Figure 65** pour la catégorie de bâtiment de la Tour Eiffel. Les **Figures 66** et **67** présentent les données prises en compte en tant que *non-bâtiment* pour l'entraînement de ce même classifieur SVM.

Il est à noter que dans la **Figure 67**, les descripteurs attribués à la classe *non-bâtiment* sont ceux de la catégorie *bâtiment* correspondant aux autres catégories. Tous les descripteurs des autres images ne sont donc pas pris en compte pour cet entraînement, mais uniquement les points clés extraits des monuments représentés dans la base de données.

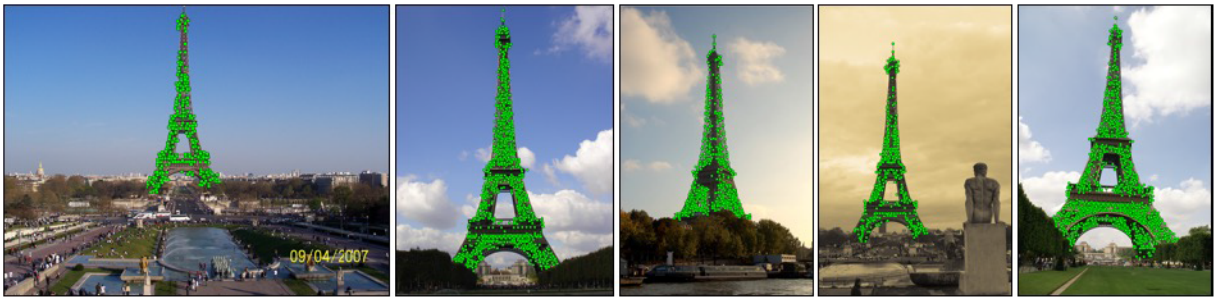


Figure 65. Descripteurs donnés de la classe *bâtiment* pour l'entraînement du modèle SVM adapté à la catégorie de bâtiment de la Tour Eiffel.

Figure 66. Première partie des descripteurs donnés de la classe *non-bâtiment* pour l'entraînement du modèle SVM adapté à la catégorie de bâtiment de la Tour Eiffel prenant en compte uniquement les images de ladite catégorie.

Figure 67. Deuxième partie des descripteurs donnés de la classe *non-bâtiment* pour l'entraînement du modèle SVM adapté à la catégorie de bâtiment de la Tour Eiffel prenant en compte les images des autres catégories de bâtiments.

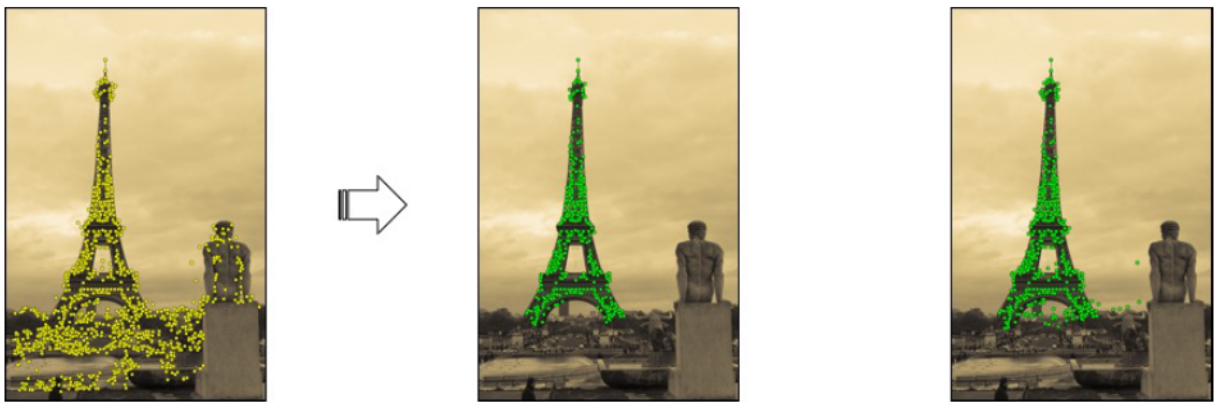
Cette méthode présente l'avantage que chaque catégorie de bâtiment est représentée par un classifieur adapté et entraîné au plus proche du dit bâtiment.

Analysons à présent la pertinence d'utiliser un modèle de classification multiple par rapport à une approche de classification globale, en nous mettant dans le cadre idéal où le classifieur optimal est connu a priori pour chaque image à traiter.

V.2. Classification multiple versus classification globale

Nous exposons dans cette section les résultats de la classification des descripteurs locaux sur différents bâtiments de la base de données Paris6k. Les **Figures 68** à **73** illustrent les résultats obtenus avec des classifieurs SVM entraînés spécifiquement pour la catégorie de bâtiments dédiée. Dans chacune de ces figures sont exposées les prédictions d'attribution des descripteurs pour une des 5 images prise en compte pour l'entraînement du modèle SVM et une image de la même catégorie retenue uniquement pour la phase de test. En parallèle sont rappelés les résultats obtenus par le modèle de classification SVM obtenu par un entraînement global.

Dans ce paragraphe, nous présentons uniquement les résultats obtenus dans le cas optimal où le classifieur SVM est correctement choisi pour la catégorie de bâtiment donnée, afin d'évaluer de manière indépendante la capacité de ces classifieurs à discriminer les points *bâtiment* de ceux *non bâtiment*.



a) **b)** **c)**

Figure 68. Résultats de la classification de descripteurs SIFT sur une image d'entraînement de la catégorie Tour Eiffel.
a) Ensemble de points d'intérêt initiaux **b)** Ensemble des points d'intérêt filtrés avec un classifieur SVM adapté à ce bâtiment **c)** Rappel des points d'intérêt filtrés avec un unique classifieur global (*cf.* **Chapitre IV**)

a) b) c)

Figure 69. Résultats de la classification de descripteurs SIFT sur une image de test de la catégorie Tour Eiffel.
a) Ensemble de points d'intérêt initiaux **b)** Ensemble des points d'intérêt filtrés avec un classifieur SVM adapté à ce bâtiment **c)** Rappel des points d'intérêt filtrés avec un unique classifieur global (*cf.* **Chapitre IV**)

a) b) c)

Figure 70. Résultats de la classification de descripteurs SIFT sur une image d'entraînement de la catégorie Notre Dame.
a) Ensemble de points d'intérêt initiaux **b)** Ensemble des points d'intérêt filtrés avec un classifieur SVM adapté à ce bâtiment **c)** Rappel des points d'intérêt filtrés avec un unique classifieur global (*cf.* **Chapitre IV**)

a) b) c)

Figure 71. Résultats de la classification de descripteurs SIFT sur une image de test de la catégorie Notre Dame.
a) Ensemble de points d'intérêt initiaux **b)** Ensemble des points d'intérêt filtrés avec un classifieur SVM adapté à ce bâtiment **c)** Rappel des points d'intérêt filtrés avec un unique classifieur global (*cf.* **Chapitre IV**)

a) b) c)

Figure 72. Résultats de la classification de descripteurs SIFT sur une image d'entraînement de la catégorie Sacré Cœur.
a) Ensemble de points d'intérêt initiaux **b)** Ensemble des points d'intérêt filtrés avec un classifieur SVM adapté à ce bâtiment **c)** Rappel des points d'intérêt filtrés avec un unique classifieur global (*cf.* **Chapitre IV**)

a) **b)** **c)**

Figure 73. Résultats de la classification de descripteurs SIFT sur une image de test de la catégorie Sacré Cœur.
a) Ensemble de points d'intérêt initiaux **b)** Ensemble des points d'intérêt filtrés avec un classifieur SVM adapté à ce bâtiment **c)** Rappel des points d'intérêt filtrés avec un unique classifieur global (*cf.* **Chapitre IV**)

Par comparaison avec les résultats obtenus avec l'unique classifieur SVM global introduit au **Chapitre IV**, nous pouvons observer que le nombre de points attribués à la classe *bâtiment* diminue fortement. En particulier, aucun point d'intérêt en dehors du bâtiment considéré (*i.e.*, faux positif) n'est attribué à cette classe sur les images d'entraînement, autrement dit, nous n'observons aucun faux positif avec les images d'entraînement.

Bien que ce résultat soit attendu, il est à noter qu'avec un unique classifieur global pour l'ensemble des bâtiments sans distinction, les bâtiments environnants peuvent être retenus comme faisant parti de la classe *bâtiment* bien qu'ils ne correspondent à aucun des monuments clés recherchés. Dans ce cas, le mur au premier plan compte de nombreux points d'intérêts marqué *bâtiment* alors que dans le cas d'une classification multiple adaptée à chaque bâtiment, seuls les points extraits qui correspondent effectivement au monument d'intérêt sont correctement attribués à la classe *bâtiment*. Cela montre bien l'avantage réel à entraîner séparément un classifieur SVM adapté à chaque catégorie de monuments définis dans la base de données.

D'autre part, après généralisation aux images de test non prises en compte pour l'entrainement, nous pouvons remarquer aussi une nette diminution du nombre de faux positifs dans les prédictions des descripteurs attribués à la classe *bâtiment*.

Le **Tableau 4** présente le nombre de faux positifs recensés pour chaque catégorie par rapport au nombre total de descripteurs extraits dans chaque image dénombrés dans la vérité terrain de la catégorie donnée pour la base de données Paris6k. Ces résultats sont à comparer à ceux obtenus avec une classification globale **Tableau 5**. Nous pouvons observer que nous réduisons grandement le nombre de faux positifs obtenus avec notre méthode de classification multiple.

Catégories	Faux positifs par image	Descripteurs *bâtiment* par image	Ratio
La Défense	1,660	93,855	0,018
Tour Eiffel	15,240	166,647	0,091
Invalides	16,100	235,869	0,068
Louvre	5,500	467,868	0,012
Moulin Rouge	9,700	157,354	0,062
Musée d'Orsay	139,660	594,306	0,235
Notre Dame	45,360	992,118	0,046
Panthéon	8,140	440,794	0,018
Pompidou	90,560	2593,882	0,035
Sacré Cœur	22,380	576,483	0,039
Arc de Triomphe	30,440	400,637	0,076
Moyenne globale	34,976	610,892	0,064

Tableau 4. Rapport du nombre moyen de faux positifs par image pour la classe *bâtiment* par image en fonction du nombre total moyen par image de descripteurs attribués à la classe *bâtiment* dans la base de données Paris6k dans le cas d'une classification entraînée de façon adaptée pour chaque catégorie de bâtiment de la base de données.

Catégories	Faux positifs par image	Descripteurs *bâtiment* par image	Ratio
La Défense	372,950	612,100	0,609
Tour Eiffel	425,050	1040,950	0,408
Invalides	457,950	1027,150	0,446
Louvre	560,250	1227,200	0,457
Moulin Rouge	517,850	1066,400	0,486
Musée d'Orsay	1198,400	1768,900	0,677
Notre Dame	249,850	2013,350	0,124
Panthéon	200,100	1114,550	0,180
Pompidou	383,400	3025,300	0,127
Sacré Cœur	285,450	1243,400	0,230
Arc de Triomphe	464,450	1295,250	0,359
Moyenne globale	465,064	1403,141	0,373

Tableau 5. Rapport du nombre moyen de faux positifs par image pour la classe *bâtiment* par image en fonction du nombre total moyen par image de descripteurs attribués à la classe *bâtiment* dans la base de données Paris6k dans le cas d'une classification entraînée de façon adaptée pour chaque catégorie de bâtiment de la base de données.

De la même manière, le nombre de faux positifs est significativement réduit lorsque le classifieur adapté est utilisé.

Les images comptant le plus de prédictions faux positifs s'expliquent par le fait qu'elles représentent des images panoramiques avec de nombreux bâtiments présents dans le voisinage du monument d'intérêt. De ce fait, bien que chaque classifieur soit spécifiquement entraîné pour chaque catégorie de la vérité terrain, cette phase d'entraînement ne compte en tout uniquement que les 55 images définies dans la vérité terrain. Se focalisant sur les bâtiments recherchés, les objets perturbateurs sont principalement des piétons, des véhicules, de la végétation et simplement quelques bâtiments voisins. De ce fait, une vue panoramique contient de nombreux bâtiments non pris en compte pendant l'entraînement du classifieur qui peuvent par la suite présenter des similarités avec le bâtiment recherché et donc attribuer des descripteurs extraits des bâtiments voisins à la classe *bâtiment*. Ce type d'images correspond aux objets perturbateurs les plus difficiles à écarter.

Cependant, ces cas restent marginaux et cette solution de filtrage par classifieur SVM adapté à chaque catégorie d'image présente une amélioration significative.

La problématique principale qui reste à résoudre est de pouvoir sélectionner le classifieur correspondant effectivement à la catégorie de bâtiment adaptée. Lors de la phase de test, nous n'avons aucune connaissance *a priori* concernant la catégorie de bâtiment à laquelle appartient chaque image. Pour obtenir un système de classification automatique, nous présentons dans la suite différentes stratégies permettant de choisir le classifieur à mettre en œuvre afin d'isoler le bâtiment recherché du reste de l'image. Cela revient à considérer une stratégie de fusion des classifieurs indépendants.

V.3. Fusion et choix du modèle SVM adapté

Pour une image donnée, 11 classifieurs sont donc disponibles pour filtrer les descripteurs extraits de cette image. La problématique est donc de choisir le modèle SVM le plus pertinent dans une phase de fusion.

Cette fusion peut s'opérer à différents niveaux. Une première possibilité consiste à considérer un niveau local. Dans ce cas un classifieur est choisi pour chaque point d'intérêt/descripteur de l'image.

Le choix du modèle peut cependant être aussi calculé au niveau global de l'image. Nous explicitons dans les paragraphes suivants les deux approches proposées.

V.3.1. Métrique d'évaluation locale par descripteur

Dans ce cas, chaque descripteur d'une image donnée est attribué à la classe *bâtiment* ou à la classe *non-bâtiment* suivant les différents modèles SVM entraînés. Nous choisissons de sélectionner le modèle classifiant le descripteur selon la plus grande probabilité d'appartenance à une catégorie donnée. La mesure de probabilité considérée est inspirée des travaux présentés dans [Plat99] et est définie dans l'équation (25) :

$$P(X) = \frac{1}{1 + \exp\left(-|\Delta_X|\right)} \tag{25}$$

où X et Δ_X désignent respectivement le descripteur associé au point d'intérêt considéré et la valeur absolue de la distance signée du point X à la marge du classifieur SVM considéré.

Notons que la probabilité $P(X)$ sera toujours inférieure à 1. Lorsque le point considéré est loin de l'hyperplan frontière, $P(X)$ approche la valeur 1, ce qui signifie que la classe attribuée est certaine. Au contraire, pour des points situés proches de la marge du classifieur SVM, la distance Δ_X est proche de 0. Dans ce cas, la probabilité $P(X)$ tend vers la valeur de 0,5, ce qui exprime l'ambiguïté de la prédiction.

Chaque point d'intérêt fournit donc une prédiction binaire de classe (*bâtiment* ou *non bâtiment*) et un score de probabilité. Cela permet par la suite de filtrer l'ensemble des descripteurs d'une image donnée, en ne conservant que ceux de la classe *bâtiment*.

L'exemple illustré **Figure 74** présente les différentes prédictions possibles pour un point d'intérêt extrait d'une image donnée suivant les différents modèles SVM entraînés. Chaque classifieur permet d'attribuer les points d'intérêt considérés à la classe *bâtiment* ou *non-bâtiment*, avec un score de probabilité associé.

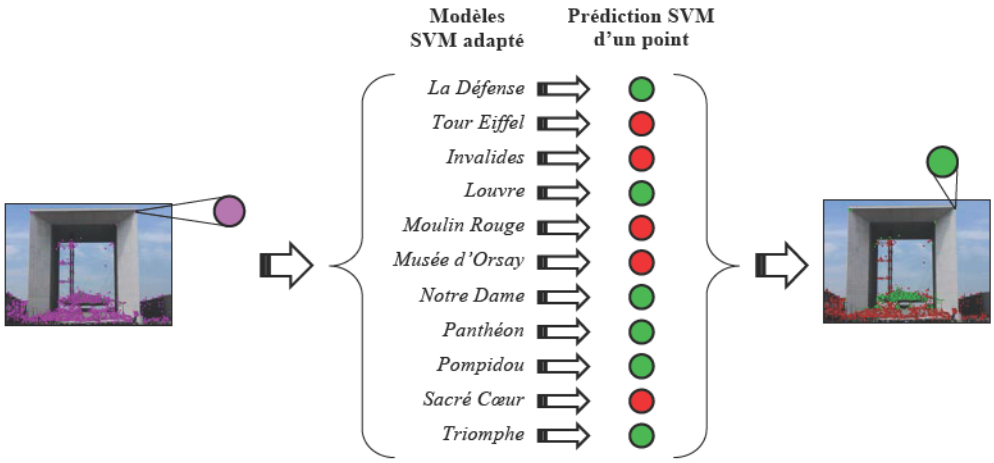


Figure 74. Exemple de prédiction SVM point clé par point clé au niveau local d'une image de la base de données Paris6k.

Finalement, la fusion des différents modèles SVM entraînés est effectuée au niveau local. Pour chaque point d'intérêt extrait d'une image, les différents modèles SVM entraînés précédemment sont appliqués pour prédire la classe du point donné (**Figure 74**). Le résultat de classification final est obtenu en prenant en compte les probabilités de prédiction d'attribution à l'une ou l'autre classe

(*bâtiment* ou *non-bâtiment*), sur l'ensemble des classifieurs considérés. Ainsi, le modèle qui fournit la plus haute probabilité de prédiction est sélectionné pour définir la classe à laquelle est attribué le point courant. Soulignons que le même processus est conduit de manière indépendante pour chacun des points extraits de l'image.

Cette stratégie de classification locale par point d'intérêt individuel reste néanmoins très sensible au sens où dans une même image nous constatons la présence d'un nombre élevé de catégories identifiés. Afin de s'affranchir de cet inconvénient, nous proposons une deuxième approche de fusion, fondée sur une métrique d'évaluation globale, qui permet de globaliser la décision sur l'ensemble des points d'intérêt d'une image.

V.3.2. Métrique d'évaluation globale par image

Chaque image est filtrée via les différents classifieurs entraînés (un classifieur différent est entraîné pour chaque catégorie de bâtiments de la base de données) comme illustré **Figure 75**.

Ici, un modèle SVM donné permet de classifier l'ensemble des points d'intérêt d'une image et non plus un unique descripteur. Nous obtenons donc ainsi 11 résultats de classification pour une même image dont les différents points clés sont attribués de la même façon à l'une ou l'autre des classes *bâtiment* et *non-bâtiment*.

Ces différentes prédictions ne sont donc pas toutes pertinentes et un modèle en particulier doit être sélectionné pour répartir au mieux ces descripteurs. L'objectif est alors de choisir le classifieur à utiliser de façon à isoler le bâtiment recherché de la façon la plus précise possible.

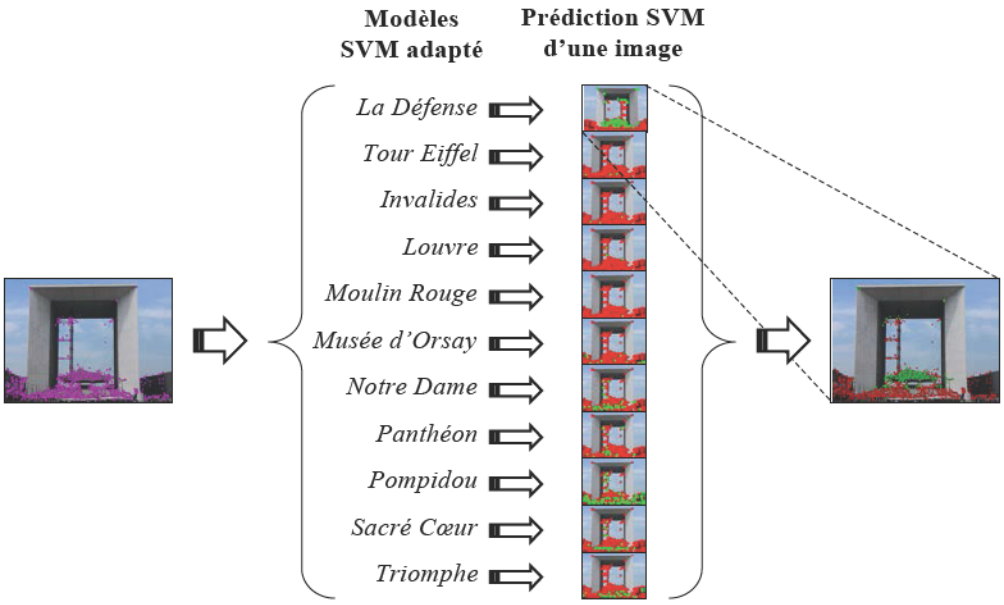


Figure 75. Exemple de prédiction SVM au niveau global pour l'ensemble des descripteurs d'une image de la base de données Paris6k.

Deux différents critères d'évaluations sont proposés afin de choisir le classifieur le plus adapté/pertinent. Un premier concerne le nombre de points clés attribués pour une image donnée à la classe *bâtiment* par chaque classifieur. Il est détaillé dans la section suivante.

V.3.2.1. Premier critère : le nombre de points clés attribués à la classe *bâtiment*

L'objectif est de conserver uniquement les points attribués à la classe *bâtiment* pour la construction d'un vocabulaire plus pertinent. Cette étape de filtrage réduisant le nombre de points d'intérêt, intuitivement nous pouvons penser que les classifieurs ne correspondant pas à la catégorie adaptée à l'image traitée retireront quasiment tous les points clés alors que le classifieur correct conserve un nombre maximum de descripteurs du bâtiment recherché, comme illustré **Figure 76**. Ainsi, nous retenons le classifieur conservant le nombre maximal de points identifiés comme appartenant à la catégorie *bâtiment*.

a) **b)**

Figure 76. Exemples de deux résultats de prédiction SVM avec deux classifieurs différents pour une image donnée : **a)** avec un classifieur correspondant à la bonne catégorie (ici *Tour Eiffel*) ; **b)** avec un classifieur formé avec une catégorie différente.

Cette solution semble à première vue correspondre à notre problématique de sélection du classifieur le plus adapté. Cependant, une difficulté à surmonter est liée au nombre de points d'intérêt extraits par image, qui peut varier significativement entre différentes catégories, comme explicité dans le **Tableau 6**.

Par exemple, la catégorie d'images représentant le centre Pompidou dans Paris6k compte 3,5 fois plus de descripteurs que celle représentant L'Arche de la Défense. Par conséquent, même en réduisant les points clés conservés par filtrage SVM en ne prenant en compte que la classe *bâtiment*, cette réduction reste relative à chaque catégorie. Si initialement les images du centre Pompidou comptent plus de descripteurs que les images de l'Arche de la Défense, le classifieur entraîné avec les images du centre Pompidou comptent plus de données d'entrée dans la phase d'entraînement que le classifieur entraîné pour les images de l'arche de la Défense.

Le nombre de points clés conservés après filtrage est présenté pour chaque classifieur entraîné spécifiquement avec chaque catégorie dans le **Tableau 7**. Celui-ci présente le cumul des résultats de la classification des descripteurs pour toutes les images de la base de données, en prenant compte des différents filtrages possibles lors du choix du classifieur SVM le plus adapté.

Nous pouvons observer que l'écart du nombre de descripteurs conservés via deux classifieurs différent se creuse. Ainsi, en reprenant l'exemple de comparaison dans la base de données Paris6k avec les classifieurs entraînés avec les images du centre Pompidou d'une part et L'Arche de la Défense d'autre part, le ratio entre le nombre de points clés conservés sur l'ensemble de la base de données est maintenant de plus de 50. Ce phénomène s'explique par le fait que les images représentant le centre Pompidou sont plus centrées sur le monument lui-même qui se distingue très fortement de l'architecture des autres bâtiments. Cependant, cela signifie aussi que le classifieur entraîné avec les images du centre Pompidou est choisi pour classifier les points *bâtiment* et *non-bâtiment* pour des images non adaptées, simplement à cause du plus grand nombre de descripteurs isolés par ce classifieur et ce indépendamment de la catégorie de bâtiments considérée.

Paris6k	
Catégories	**Descripteurs par image**
La Défense	1 372
Tour Eiffel	1 989
Invalides	2 499
Louvre	2 507
Moulin Rouge	2 465
Musée d'Orsay	4 414
Notre Dame	3 529
Panthéon	1 838
Pompidou	4 756
Sacré Cœur	2 736
Arc de Triomphe	2 534
Moyenne	**2 528**

Oxford5k	
Catégories	**Descripteurs par image**
All Souls	3 317
Ashmolean	2 018
Balliol	3 125
Bodleian	2 979
Christ Church	2 016
Cornmarket	3 660
Hertford	2 991
Keble	3 305
Magdalen	2 482
Pitt Rivers	3 335
Radcliffe Camera	3 092
Moyenne	**2 871**

Tableau 6. Nombre moyen de points d'intérêt extraits par image pour chaque catégorie de bâtiment dans la vérité terrain des bases de données Paris6k et Oxford5k.

Paris6k		Oxford5k	
Classifieur	**Descripteurs retenus après filtrage**	**Classifieur**	**Descripteurs retenus après filtrage**
La Défense	6	All Souls	128
Tour Eiffel	40	Ashmolean	65
Invalides	38	Balliol	62
Louvre	55	Bodleian	77
Moulin Rouge	21	Christ Church	26
Musée d'Orsay	169	Cornmarket	167
Notre Dame	129	Hertford	132
Panthéon	53	Keble	234
Pompidou	368	Magdalen	39
Sacré Cœur	88	Pitt Rivers	35
Arc de Triomphe	82	Radcliffe Camera	135
Moyenne	**95**	**Moyenne**	**100**

Tableau 7. Nombre moyen par image de descripteurs conservés après filtrage SVM en fonction du classifieur choisi pour l'ensemble des images des bases de données Paris6k et Oxford5k.

D'autre part, pour chaque point clé, le classifieur SVM renvoie une valeur signée de la distance du point donné à l'hyperplan calculé. Afin d'exploiter cette distance dans une optique globale de pondération, une seconde stratégie est proposée.

V.3.2.2. Deuxième critère : la probabilité attribuée à la prédiction SVM

Cette seconde approche prend en compte la distance à la marge donnée par les prédictions SVM. En effet, plus la valeur absolue de cette distance est importante, plus la classification est certifiée. *A contrario*, plus le point est proche de la marge de séparation, plus la prédiction est indécise. Le principe ici est de considérer la vraisemblance issue de cette distance prédiction. Nous généralisons ce score de confiance pour une image entière en considérant la valeur moyenne des scores de confiance $P(X)$ sur l'ensemble des points attribués à la classe *bâtiment* par le classifieur SVM. Le classifieur conduisant à la probabilité la plus élevée sera retenu pour une image donnée.

Cette stratégie permet de prendre en compte pas seulement le nombre de points d'intérêt attribués à la classe *bâtiment*, mais également une notion de *certitude*. Sur l'ensemble des bases de données Paris6k et Oxford5k, les scores de confiance sont calculés pour l'ensemble des images de la base de données considérée avec chacun des différents classifieurs SVM entraînés.

Ces scores sont présentés dans le **Tableau 8**. Ils nous permettent d'évaluer l'influence de chacun des modèles SVM entraînés et engendrés précédemment. La disparité des scores est ici réduite comme le montre l'écart type moyen calculé entre les différentes catégories sur l'ensemble de la base de données dans le **Tableau 8**. Cela permet d'éviter la prépondérance d'un modèle SVM d'une catégorie de

bâtiment particulière sur une autre (à cause d'un surnombre de descripteurs détectés initialement par exemple).

Paris6k	
Classifieur	**Score de confiance**
La Défense	0,526
Tour Eiffel	0,541
Invalides	0,542
Louvre	0,531
Moulin Rouge	0,539
Musée d'Orsay	0,570
Notre Dame	0,524
Panthéon	0,560
Pompidou	0,453
Sacré Cœur	0,572
Arc de Triomphe	0,561
Moyenne	**0,538**
Écart type	**0,031**

Oxford5k	
Classifieur	**Score de confiance**
All Souls	0,570
Ashmolean	0,560
Balliol	0,562
Bodleian	0,569
Christ Church	0,552
Cornmarket	0,571
Hertford	0,569
Keble	0,577
Magdalen	0,556
Pitt Rivers	0,553
Radcliffe Camera	0,565
Moyenne	**0,564**
Écart type	**0,008**

Tableau 8. Moyenne et écart type des scores de confiance pour l'ensemble des images des bases de données Paris6k et Oxford5k, en fonction des différentes catégories d'images.

Etudions à présent les performances des deux approches de fusion sur la base des résultats expérimentaux obtenus sur les vérités terrain disponibles pour les bases de données Paris6k et Oxford5k.

V.4. Choix du classifieur optimal sur la vérité terrain : résultats expérimentaux

Les **Figures 77** et **78** présentent les classifieurs choisis pour chaque image de la vérité terrain suivant le premier critère, *i.e.* le nombre de points clés conservés après filtrage respectivement pour les bases de données Paris6k et Oxford5k.

En moyenne, le classifieur adapté est correctement choisi à 54,41% (entre 19% pour la catégorie la Défense et 100% pour la catégorie du musée de Pompidou) pour la base de données Paris6k et 87,82% (entre 28% pour la catégorie Magdalen et 100% pour les catégories Bodleian, Cornmarket, Keble et Pitt Rivers) pour la base données Oxford5k.

Les **Figures 79** et **80** présentent les classifieurs sélectionnés pour chaque image de la vérité terrain en appliquant le deuxième critère, *i.e.*, le score de confiance globalisé des prédictions SVM respectivement pour les bases de données Paris6k et Oxford5k. En moyenne, le classifieur adapté est correctement choisi à 66,32% (entre 35% pour les catégories Moulin Rouge et du musée d'Orsay et 100% pour la catégorie du musée de Pompidou) pour la base de données Paris6k et 87,49% (entre 46% pour la catégorie Magdalen et 100% pour les catégories Balliol, Bodleian, Keble et Pitt Rivers) pour la base de données Oxford5k.

En ce qui concerne le premier critère retenu, il souffre des limitations déjà évoquées (*cf.* **Paragraphe V.3.2.1**), liées aux catégories d'images présentant un nombre relativement important de points d'intérêt, comme c'est le cas de la classe *Pompidou* pour la base Paris6k. Ainsi, cette catégorie se retrouve détectée comme prépondérante à plus de 40% des cas pour 7 des 11 catégories de cette base.

Ces erreurs sont toutefois fortement atténuées lorsqu'on considère le deuxième critère de sélection. Dans ce cas, chaque catégorie de bâtiment choisit en majorité le classifieur adapté.

Cette évolution n'est pas aussi flagrante pour la base de données Oxford5k. Cela peut s'expliquer par le fait que cette base compte beaucoup moins d'images panoramiques qui ne sont, de plus, pas comptabilisées dans les images de la vérité terrain.

Les critères présentés ici sont donc bien à prendre en compte dans le choix du classifieur SVM adapté au filtrage des points clés.

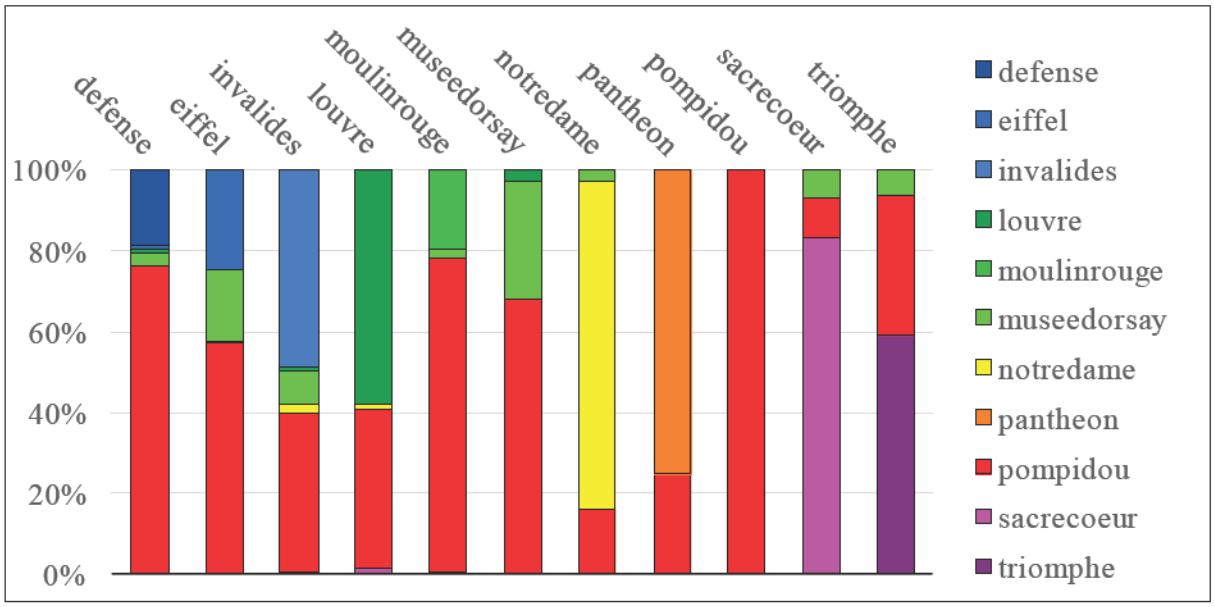


Figure 77. Pourcentage de choix de classifieur suivant le premier critère prenant en compte le nombre de points clés conservés après filtrage pour les images de la vérité terrain de la base de données Paris6k.

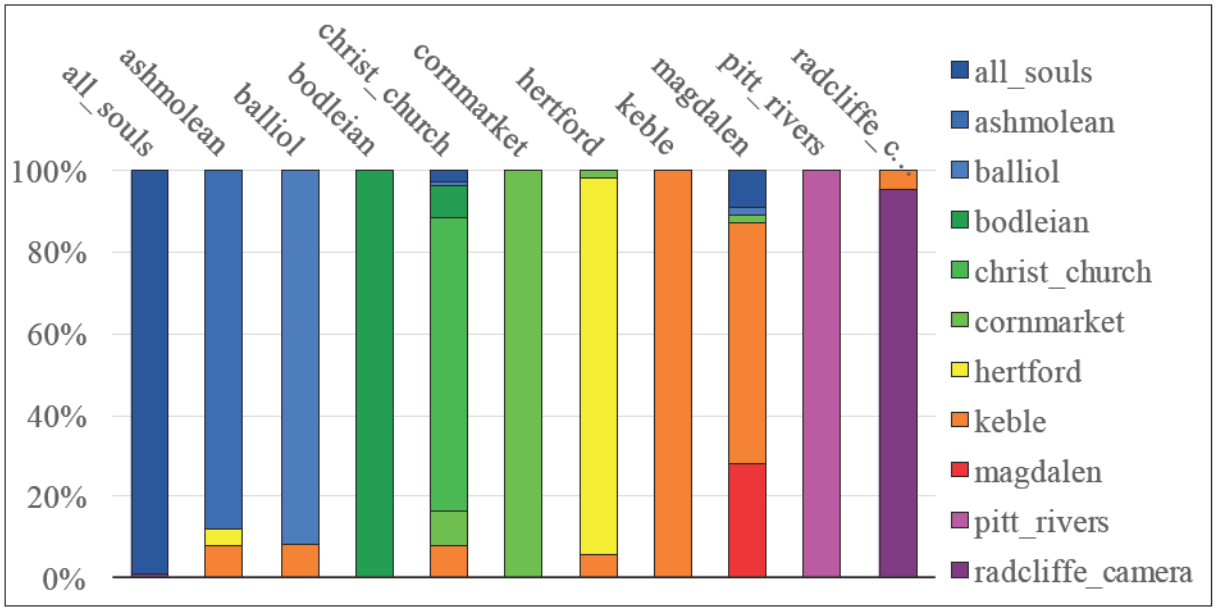


Figure 78. Pourcentage de choix de classifieur suivant le premier critère prenant en compte le nombre de points clés conservés après filtrage pour les images de la vérité terrain de la base de données Oxford5k.

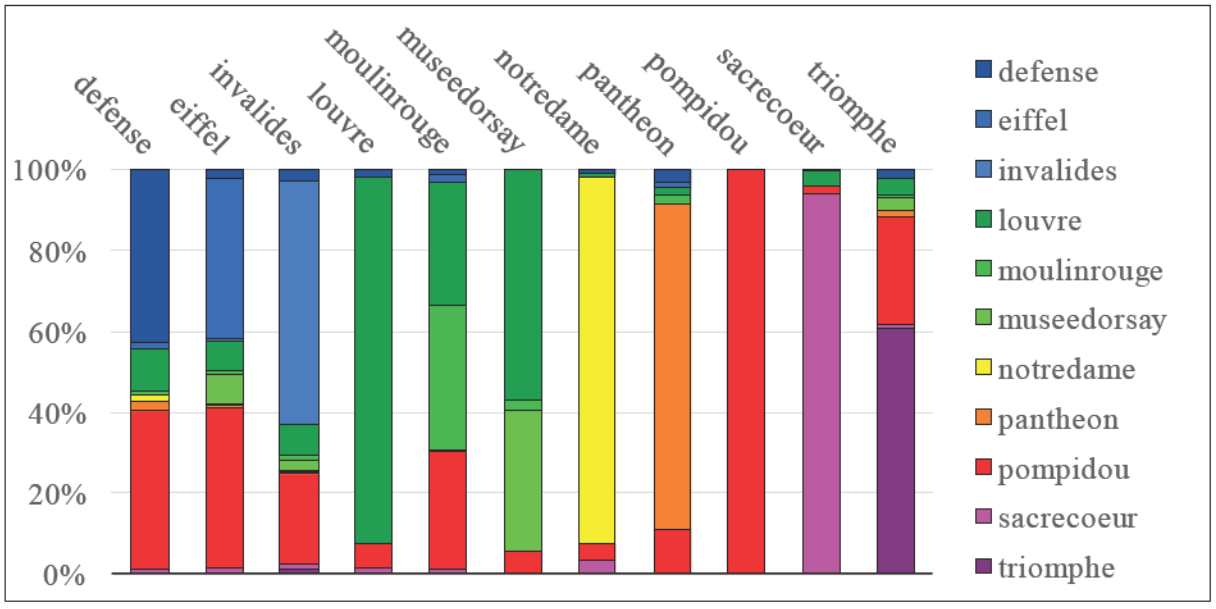


Figure 79. Pourcentage de choix de classifieur suivant le deuxième critère prenant en compte le score de confiance de filtrage pour les images de la vérité terrain de la base de données Paris6k.

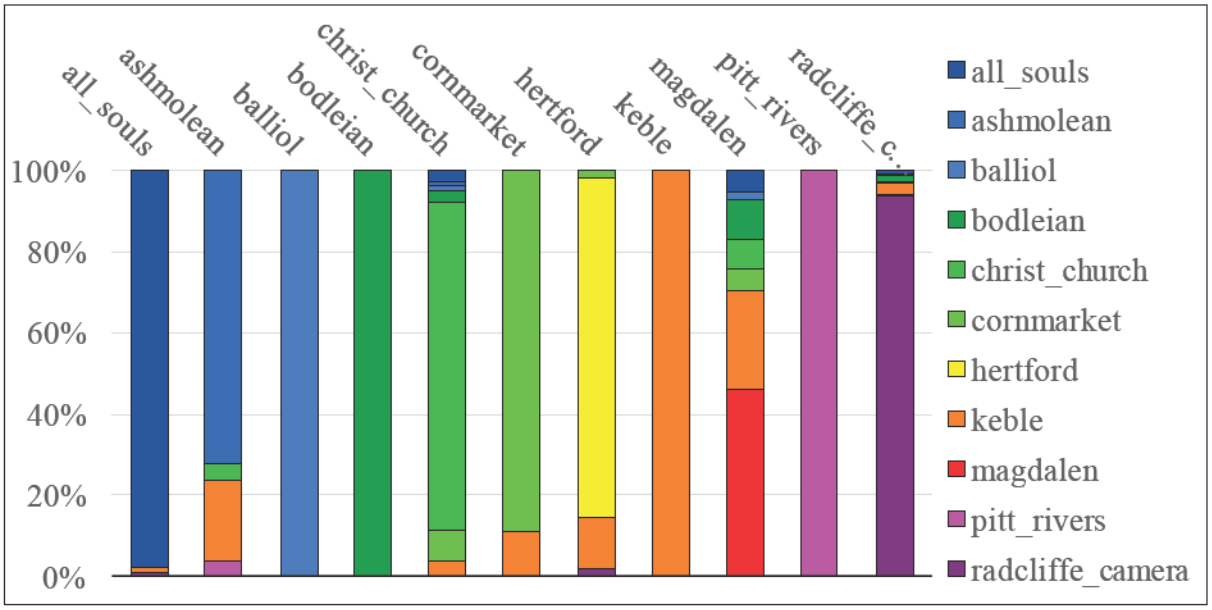


Figure 80. Pourcentage de choix de classifieur suivant le deuxième critère prenant en compte le score de confiance de filtrage pour les images de la vérité terrain de la base de données Oxford5k.

Pour la base de données Oxford5k, les résultats de classification sémantique sont similaires selon les deux critères de sélection mais pour la base de données Paris6k, le second critère de classification sémantique est plus pertinent. Les images de la vérité de terrain étant moins cohérentes pour chaque catégorie de bâtiments dans la base de données Paris6k, cela explique les résultats de classification moins performants que pour la base de données Oxford5k.

Par exemple, dans la base de données Paris6k pour la catégorie de bâtiments de la classe La Défense, les images utilisées en requêtes présentent l'arche de La Défense **Figure 81** alors qu'il n'est pas rare que l'arche ne soit que très peu visible dans les images de la vérité de terrain comme illustré avec quelques exemples **Figure 82**.

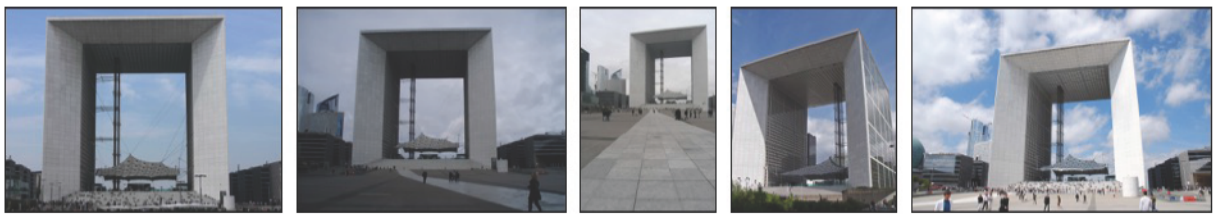


Figure 81. Images requêtes de la catégorie d'images représentant l'arche de La Défense dans la base de données Paris6k.

Figure 82. Exemples d'images de la vérité terrain issue de la catégorie d'images représentant l'arche de La Défense dans la base de données Paris6k.

V.5. Bilan

Dans ce chapitre, nous avons montré qu'un classifieur entraîné spécifiquement pour une catégorie de bâtiment conduit à une sélection de points clés plus adaptés et ainsi plus précise. Cette constatation semble évidente mais la problématique sous-jacente qui émerge concerne la sélection automatique du classifieur le plus adapté, pour chaque image considérée et sans aucune connaissance *a priori* de sa catégorie. Nous avons donc défini deux critères de sélection permettant d'isoler le bâtiment recherché sur chaque image de façon automatique et sans connaissance a priori. Un premier s'appuie sur le nombre de points d'intérêt catégorisé dans une des catégories de la vérité terrain. Le second prend en compte la vraisemblance d'appartenance moyenne à une des catégories existantes.

Soulignons également que le modèle présenté permet par la suite de réduire le nombre de descripteurs pris en compte pour construire le vocabulaire de mots visuels, de façon à ne conserver que les mots décrivant les monuments d'intérêt.

Evoquons aussi que cette approche permet déjà de catégoriser automatiquement une image non connue dans une des catégories de la base de données. En effet, lors de la sélection du classifieur, notre stratégie est de choisir le classifieur le plus adapté. Ce classifieur est considéré comme étant correctement déterminé s'il a effectivement été entraîné avec les images de la catégorie de l'image considérée. Ainsi, la connaissance du classifieur choisi permet de déterminer la catégorie des images utilisées en entraînement et ainsi désigner la catégorie de l'image prise en requête.

L'utilisation de cet algorithme est par la suite de construire un vocabulaire plus pertinent, moins dense et moins pollué par des informations non nécessaires.

Enfin, notons que l'ensemble de ces développements est réalisé de manière indépendante, dans l'espace des descripteurs SIFT, sans aucune prise en compte des relations spatiales pouvant exister entre les différents points d'intérêt identifiés. Cela peut, à l'évidence, pénaliser le processus de classification. Afin de s'affranchir de cet inconvénient, nous avons considéré un raffinement supplémentaire, qui concerne une étape de vérification et de correction de la consistance géométrique des points d'intérêt identifiés. Elle est décrite dans le chapitre suivant.

VI. VERIFICATION ET CORRECTION GEOMETRIQUE

Résumé. Nous proposons dans ce chapitre d'ajouter dans le processus de description une étape supplémentaire de vérification de la cohérence géométrique des points d'intérêt. En effet, les deux chapitres précédents proposent un traitement des données extraites d'une image au niveau local alors que l'objectif est bien d'identifier des bâtiments entiers. Les points d'intérêts extraits d'un même objet sont donc censés être attribués à la même classe lors de la prédiction par modèle SVM. Cependant, comme nous l'avons montré, plusieurs faux positifs et faux négatifs peuvent persister. Ce chapitre introduit notamment une méthode mettant en œuvre l'influence du voisinage d'un point sur l'attribution de sa classe *bâtiment* ou *non-bâtiment* de façon à conserver une cohérence géométrique après une classification locale. Nous introduisons dans un premier temps la définition du voisinage d'un point donné et le calcul de vraisemblance moyenne au sein de ce voisinage. Cela permet de vérifier et de corriger au besoin l'attribution de la classe prédite en un point. Nous montrons ainsi qu'il est possible de réduire le nombre de faux positifs et de faux négatifs et donc affiner les données pertinentes conservées lors de la construction d'un vocabulaire par la suite.

Mots clés : relation de voisinage, cohérence géométrique, quadtree, vraisemblance moyenne

VI.1. Prédictions et probabilités des modèles SVM

A l'aide de la méthode de classification introduite au chapitre précédent, il est à présent possible de catégoriser une image inconnue et d'en filtrer les points d'intérêt correspondant au monument qui y est présent. Cette classification s'appuie sur un ensemble de modèles SVM entraînés indépendamment pour chaque catégorie de la base de données. Cependant, ces modèles sont évalués dans l'espace des descripteurs, de dimension 128 pour les descripteurs SIFT retenus dans notre cas.

Rappelons que l'objectif est de regrouper les points d'intérêts représentant l'objet recherché sous un même label *bâtiment* et d'isoler/éliminer les autres points d'intérêts de la classe *non-bâtiment*. Cependant, l'hyperplan définissant la frontière entre les deux classes de données dépend des données d'entraînement en fonction des paramètres considérés. Par conséquent, la distance à la marge évaluée par le classifieur pour certains descripteurs tests peut être trop faible pour que la classe attribuée soit fiable. Ce problème est illustré **Figure 83**.

En outre, soulignons que les points attribués de manière erronée à la classe *bâtiment* ajoutent des éléments d'information indésirable pour la construction du vocabulaire dédié.

Figure 83. Exemple de points clés attribués par erreur à la classe bâtiment. Pour des raisons de lisibilité, nous ne présentons que les points clés attribués à la classe bâtiment pour cette image de la catégorie représentant la Tour Eiffel dans la base de données Paris6k. Les points mal classifiés (les faux positifs) sont ici entourés d'un cercle rouge.

Afin de corriger ces erreurs inhérentes, il est nécessaire de renforcer le processus de classification, en y insérant des éléments de spatialisation. Ainsi, le principe consiste à ne plus considérer les points d'intérêt de manière complètement indépendante, en prenant notamment en compte leur position dans le plan image. Pour des scènes incluant des bâtiments/monuments il est naturel de s'appuyer sur une hypothèse de localisation homogène et cohérente des points d'intérêt détectés.

VI.2. Correction des prédictions par un classifieur

Nous cherchons ici à donner une cohérence géométrique à la classification des caractéristiques locales effectuées précédemment. Le but est de rejeter les points d'intérêt attribués par erreur à la classe *bâtiment*.

VI.2.1. Influence du voisinage dans la correction des prédictions obtenues

Le principe mis en avant est le suivant. Pour chaque point considéré, son voisinage est déterminé et la probabilité de la classification est évaluée pour l'ensemble des points de ce voisinage.

La classification SVM retourne pour chaque point clé une distance à l'hyperplan frontière. Cette décision est une valeur signée dont le signe dépend du côté de l'espace attribué à la donnée d'entrée par rapport à l'hyperplan (positif pour des points catégorisés comme *bâtiment* et négatif dans le cas contraire). Comme nous l'avons déjà évoqué, plus la valeur absolue de cette distance absolue est élevée, plus la donnée est éloignée dans l'espace de la frontière de décision et par conséquent, plus la classification est certaine. Le score de confiance de la prédiction SVM est défini comme étant une fonction sigmoïde présentée dans l'équation (25) page 108.

Pour chaque point d'intérêt, une valeur de vraisemblance moyenne $V(p)$ sur son voisinage est calculée, comme décrit dans l'équation (26). Notons que cette valeur peut être négative ou positive, chaque point d'intérêt du voisinage influençant ce résultat à partir de son propre score de vraisemblance évalué *a priori*.

$$V(p) = \frac{\sum_{X \in Voisinage(p)} P[X] \cdot \Delta_X}{\sum_{X \in Voisinage(p)} P[X]} \qquad (26)$$

où $P[X]$ désigne la probabilité de classification correcte du point X et Δ_X la distance à la marge du point X.

Par conséquent, chaque attribution de classe d'un point donné (*bâtiment* ou *non-bâtiment*) peut être corrigée en utilisant l'influence de la prédiction de ses voisins. Ce processus est illustré **Figure 84**.

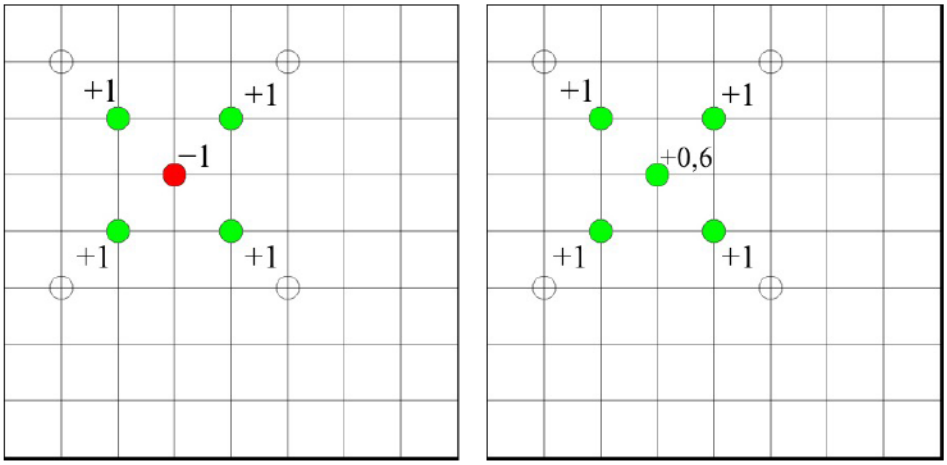


Figure 84. Exemple de correction d'un point clé appartenant à une classe erronée vers sa classe correcte suivant l'influence de la classe prédite pour ses voisins a) prédiction initiale b) prédiction corrigée.

Ici, les différents points d'intérêt sont initialement attribués à une classe *rouge* (pour les points classifiés comme *non bâtiment*) ou *verte* (pour les points *bâtiment*). Le point d'intérêt évalué pour cet exemple est attribué à la classe *rouge* quand tous ses voisins font partie de la classe *verte*. Nous pouvons alors évaluer pour chaque point de ce voisinage (le point étudié compris) sa distance à la marge de décision et donc le score de confiance de son attribution à l'une ou l'autre classe.

Dans notre exemple illustré sur la **Figure 84**, chaque point est éloigné de l'hyperplan d'une unité, *i.e.* pour chaque point X considéré dans ce voisinage, $|\Delta_X| = 1$. La prédiction émise par le classifieur SVM retourne donc la valeur signée −1 pour le point de la classe *rouge* et +1 pour les points de la classe *verte*. Dans cet exemple, le score de confiance de chaque point est donc le même, égale à $\frac{1}{1+\exp(-1)}$, c'est-à-dire 0,73. La valeur attendue pour le point d'intérêt considéré ici, en tenant compte de l'influence de son voisinage, est donc de +0,60. Le signe de la distance à la marge pour le point d'intérêt évalué ici change, par conséquent, la classe prédite pour ce point d'intérêt peut être corrigée et être attribuée à la classe *verte* grâce à l'influence de ses voisins.

Un élément essentiel dans ce processus qui reste à définir est le voisinage de chaque point d'intérêt extrait de l'image.

VI.3. Définition du voisinage d'un point d'intérêt

Nous cherchons à présent à définir le voisinage de chaque point à prendre en compte pour vérifier et corriger si besoin la classe lui ayant été attribuée initialement. Dans ce cadre, plusieurs options peuvent être prises en compte, incluant cellules rectangulaires de taille fixe ou de structures hiérarchiques arborescentes.

VI.3.1. Algorithmes de sélection des points voisins

VI.3.1.1. Voisinage rectangulaire par cellule de taille fixe

Une première solution simple est d'appliquer de considérer comme voisinage une cellule rectangulaire de taille fixe, définie *a priori* (en tenant compte, par exemple, de la taille de l'image). L'objectif est alors de déterminer, pour un point d'intérêt donné, l'ensemble des points d'intérêt intégrant la cellule qui lui est associée.

Les points d'intérêt extraits n'étant ordonnés selon aucune règle particulière, il est nécessaire d'évaluer l'ensemble des points une première fois afin de collecter les voisins du point clé dont nous effectuons la vérification. Pour chaque point extrait, celui-ci est ajouté à l'ensemble du voisinage s'il appartient à la même sous-zone définie par la cellule du point à vérifier. Ce procédé doit être effectué pour chaque point d'intérêt à vérifier.

Cette méthode est relativement simple à mettre en œuvre, cependant, elle est complexe en temps de calcul. En effet, il faut évaluer chaque point à vérifier et parcourir à chaque fois l'ensemble des points uniquement pour rechercher un voisinage. Les images de nos bases de données Paris6k et Oxfrod5k pouvant compter jusqu'à plus de 20 000 points, l'étape seule de sélection des points voisins devient complexe et chronophage.

Nous proposons donc une approche plus complexe à la mise en place mais représentant un réel avantage lors de l'évaluation.

VI.3.1.2. Voisinage par quadtree

Les arbres quaternaires (*quadtrees* dans la littérature anglophone) [FiBe74] sont une solution adaptée dans le but de trouver les plus proches voisins d'un point donné. Le principe consiste à diviser l'espace de l'image en rectangles de plus ou moins grandes tailles en fonction de la profondeur de l'arbre et des points d'intérêt présents.

Par définition, pour un arbre quaternaire chaque nœud possède tout au plus quatre enfants et assure que lorsqu'un point est ajouté à l'arbre, les nœuds sont réarrangés de telle sorte qu'aucun d'entre eux n'ait plus de quatre enfants.

Le processus d'insertion d'un point dans l'arbre se déroule comme suit. Si le sous-espace défini compte plus de quatre points d'intérêt et que le niveau maximum de l'arbre n'est pas atteint, l'espace est subdivisé de nouveau en quatre sous rectangles. Si le sous-espace défini compte moins de quatre éléments, les points sont ajoutés à l'arbre sous le nœud correspondant. Si le sous-espace compte plus de quatre éléments mais le niveau maximum de l'arbre est atteint, la feuille de l'arbre à ce niveau est alors un rectangle regroupant l'ensemble des points qu'il contient.

Chaque nœud peut être alors soit un rectangle (ayant lui-même quatre autres enfants) ou un point. La **Figure 85** illustre un exemple de construction d'un quadtree. Ce modèle accélère le processus lors de la recherche des voisins d'un point et permet de regrouper les points d'intérêt directement par régions géométriques. En effet, lors de la construction de l'arbre, les points sont regroupés en régions de tailles paramétrées définissant directement un voisinage. De ce fait, lors de la recherche du voisinage d'un point donné, ses voisins sont retrouvés directement parmi les nœuds enfants sans avoir à parcourir l'ensemble des points. De plus, une profondeur d'arbre faible (nous choisirons par la suite une profondeur de 5) suffit à obtenir un voisinage suffisamment fin.

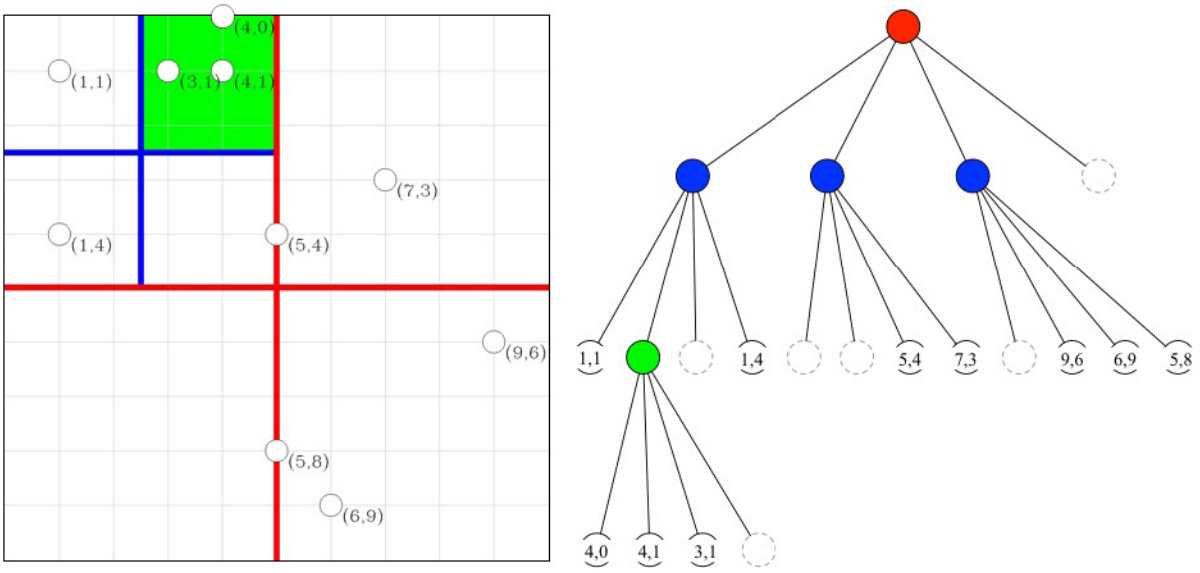


Figure 85. Exemple de construction d'un arbre quadtree. Les nœuds de couleur représentent des rectangles, et les points sont écrits avec leurs coordonnées dans le graphique. Les autres nœuds restent vides pour pouvoir accueillir de potentiels autres enfants a) représentation graphique des points et division de l'espace au cours du processus de construction du quadtree b) Représentation des points du graphique sous forme de quadtree.

Dans l'exemple illustré **Figure 85**, les voisins du point (4, 0) sont (4, 1) et (3, 1) d'après cet arbre de profondeur 3. Pour ce faire, nous partons de la racine en rouge, puis nous parcourons le premier nœud bleu puisqu'il s'agit du seul nœud de ce niveau en commun avec le point recherché et de la même manière le nœud vert est examiné. A ce stade, nous avons atteint le niveau maximum défini pour notre arbre, nous avons donc ainsi obtenu notre voisinage de trois points.

VI.3.1.3. Comparaison

Comparons maintenant les deux méthodes présentées ci-dessus en terme de complexité de calcul. Nous nous concentrons exclusivement sur la phase de recherche des points voisins. En effet, la deuxième étape étant d'évaluer l'espérance moyenne définie dans l'équation (26) pour chaque voisinage, pour un même voisinage donné le calcul est identique pour les deux approches.

Dans la méthode par cellule, aucune mise en place préalable n'est nécessaire autre que de définir la taille de chaque voisinage. Par la suite, pour chaque point A de l'image, il faut dans un premier temps obtenir les coordonnées limites de son voisinage. Dans un deuxième temps, pour chaque point B de l'image, nous vérifions si le point B est inclus dans les limites du voisinage du point A. Si tel est le cas, le point B est ajouté à la liste des voisins du point A et nous parcourons cette liste par la suite pour effectuer notre vérification dans ce voisinage ainsi défini. Dans cette phase de recherche, nous parcourons donc deux fois l'ensemble des points d'intérêt extraits de l'image. La complexité de cette algorithme ce calcul donc est en $O(n^2)$.

En ce qui concerne la deuxième approche, la construction du *quadtree* nécessite de parcourir une fois chaque point de l'image pour l'ajouter à l'arbre. La complexité de cette étape est donc en

$O(n)$. Ce processus n'est exécuté qu'une seule fois pour chaque image et fournit alors une représentation ordonnée de l'ensemble des points extraits pour image donnée.

La phase de recherche des voisins quant à elle parcourt l'arbre de façon à rejeter les branches n'aboutissant pas au point d'intérêt dont nous recherchons les points voisins. Dans cette étape l'espace est à chaque fois subdivisé de façon à ne parcourir qu'une partie de l'espace. Un nœud étant choisi à chaque étape parmi 4 au maximum, la complexité de l'étape de recherche d'un point et de son voisinage est donc en $O(\log(n))$ pour n points au total. La complexité de la procédure de recherche de voisinage pour l'ensemble des n points extraits d'une image est donc en $O(n \log(n))$ pour définir les voisinages de tous les points extraits d'une image.

La complexité de calcul de la méthode du quadtree est donc plus faible que celle de la méthode par cellule pour un grand nombre de points clés (ce qui est notre cas ici, l'ensemble des images des bases de données Paris6k etOxford5k comptant en moyenne 2 000 points clés). Nous illustrons ce fait par l'étude du temps de calcul de chacune de ses méthodes en fonction du nombre d'images à traiter (et par conséquent en fonction du nombre de points clés à évaluer). Avec peu d'images, les temps de calcul des deux méthodes sont équivalents (**Figure 86**) mais l'évolution de la différence de temps de calcul entre les deux méthodes tend à croître de façon polynomiale comme le montre la **Figure 87**. Plus le nombre d'images à évaluer est important, plus la méthode des *quadtree* est adaptée à la définition d'un voisinage pour chaque point d'une image. En effet, pour une centaine d'images, le temps de traitement avec les méthodes de voisinages rectangulaires et de quadtree est de l'ordre de la minute alors que pour l'ensemble de la base de données Paris6k (comptant 6412 images), la méthode de voisinages rectangulaires est plus lente de plus de 10 minutes par rapport à la méthode des arbres de recherche.

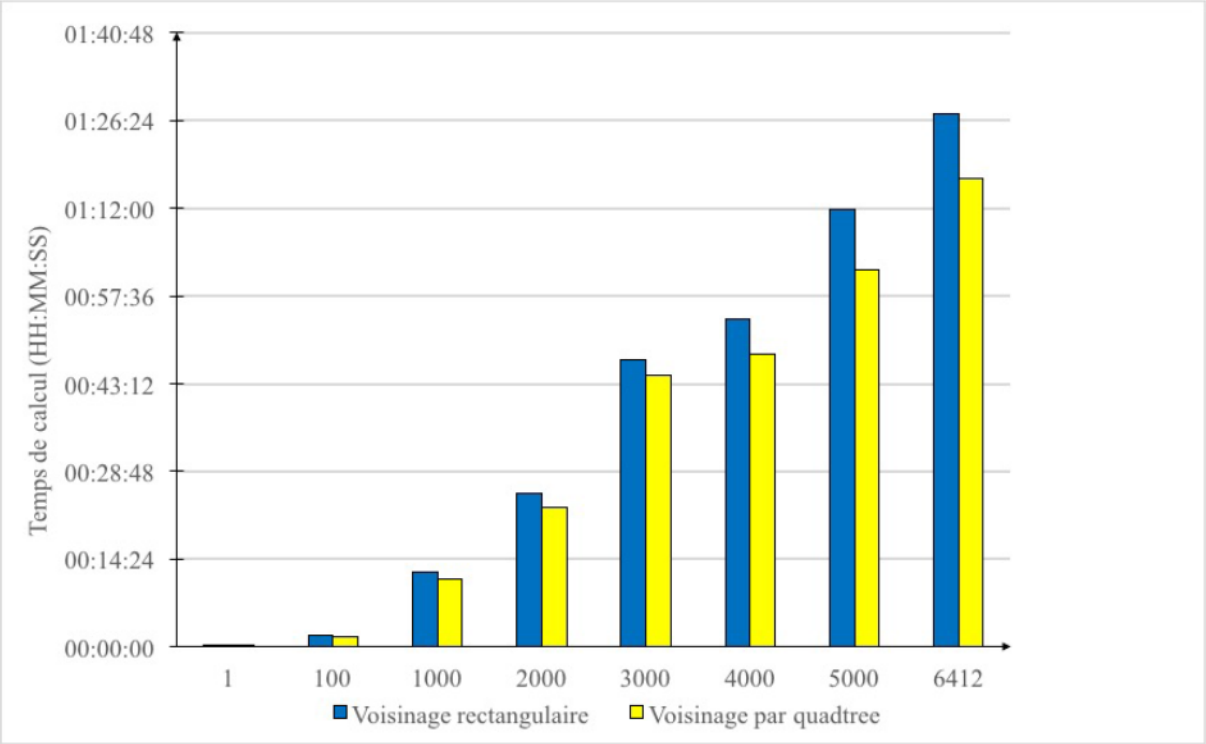


Figure 86. Différents de temps de calcul avec les méthodes par cellule et par *quadtree* en fonction du nombre d'images traitées (sur la base de données Paris6k).

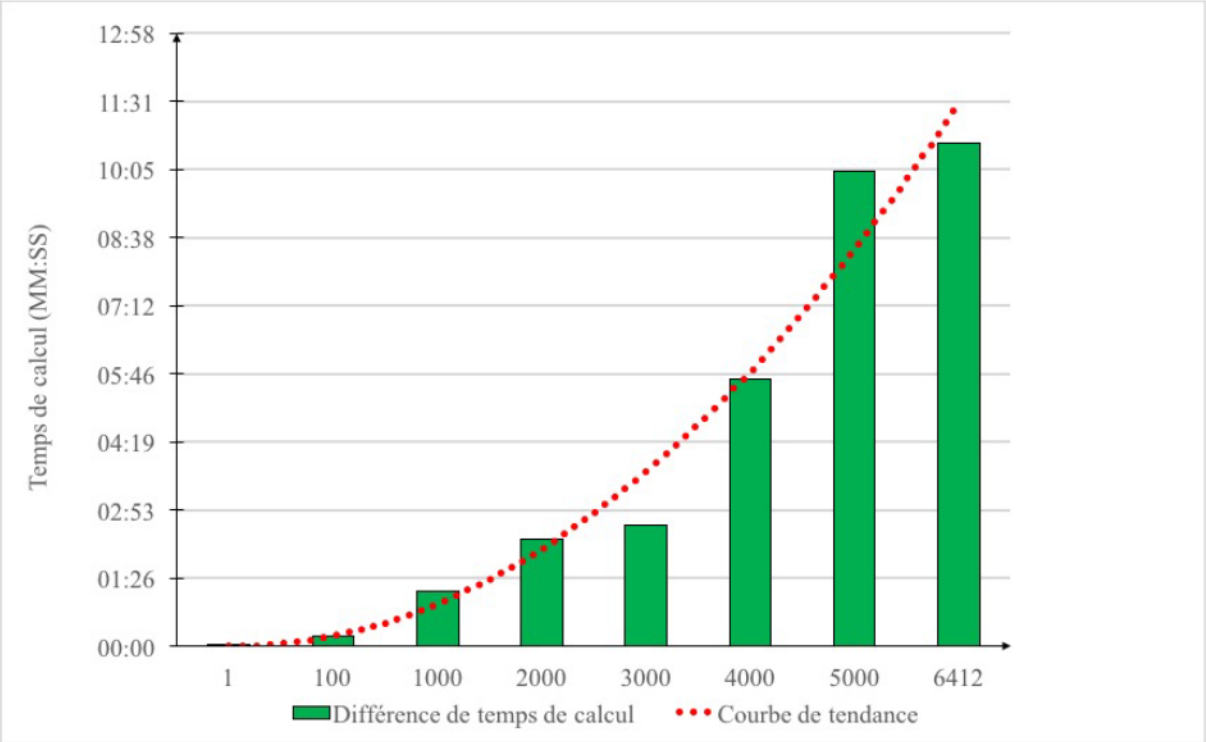


Figure 87. Différence de temps de calcul entre les méthodes par cellule et par *quadtree* en fonction du nombre d'images traitées.

VI.3.2. Choix de la taille du voisinage

Pour la première méthode faisant intervenir un voisinage de type cellule, la définition de la taille de ce voisinage est triviale. Celle-ci est directement spécifiée comme paramètre.

Dans le cas de l'utilisation d'un *quadtree*, la profondeur de l'arbre donnée en paramètre influence directement la taille de ces régions. Cet incidence du niveau maximum paramétré pour un quadtree est illustré **Figure 88**.

Nous pouvons remarquer que dans le premier cas, l'image n'est divisé qu'une seule fois. Cette configuration est visuellement semblable à une simple grille rectangulaire. Cependant, au moment de la recherche des plus proches voisins d'un point donné, tous les points de l'image ne sont pas évalués mais l'arbre défini au préalable permet de limiter le parcours des données en ne considérant que les branches auxquelles le point recherché appartient. De ce fait, en atteignant la feuille de l'arbre comptant notre point d'intérêt, nous pouvons obtenir directement ses voisins.

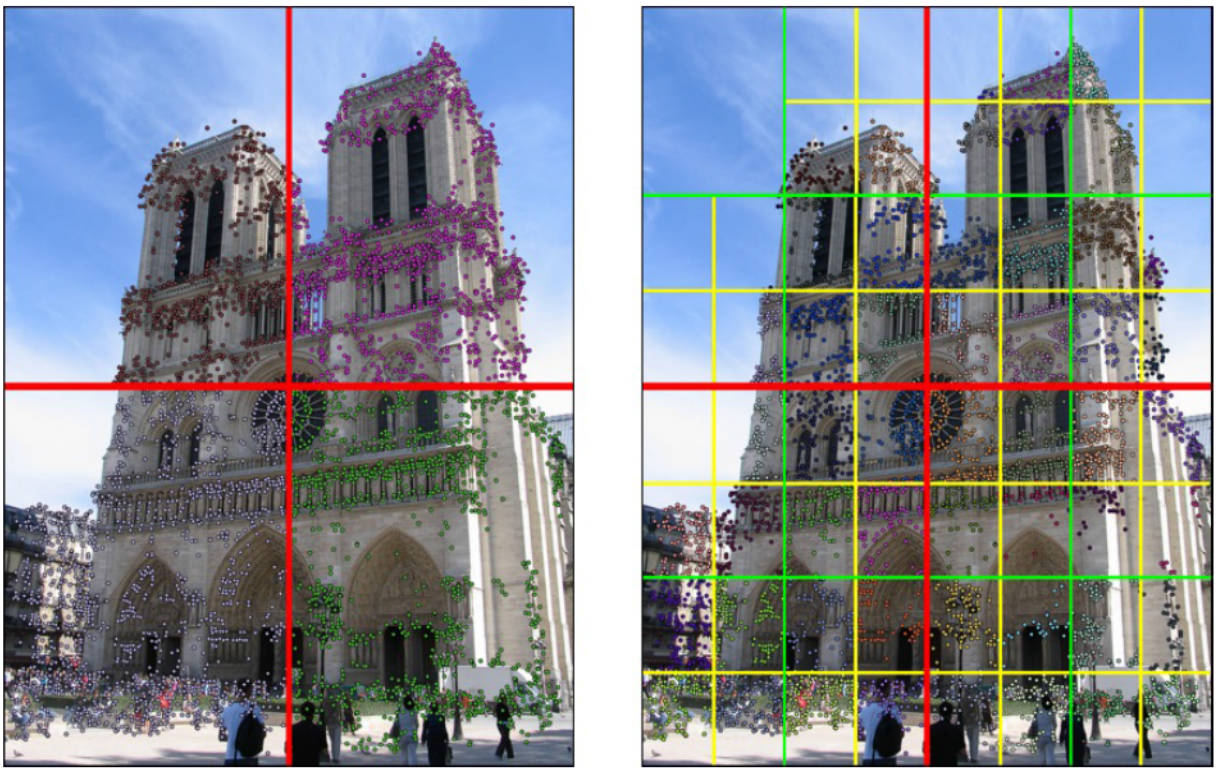


Figure 88. Clusters définis lors du processus de construction de quadtree a) le niveau maximum du quadtree est de 1 b) le niveau maximum du quadtree est de 3.

Dans le deuxième cas illustré **Figure 88**, la profondeur de notre arbre est de 3. Chaque zone est donc subdivisée au plus 3 fois. Dans le pire des cas, l'image est donc fractionnée en $2^3 \times 2^3$, c'est-à-dire 64 zones. Cependant, dans notre exemple il est a noté que la première zone délimitée par un rectangle vert en haut à gauche n'est pas subdivisée. En effet, lors de la construction de notre *quadtree*, ce nœud ne comporte que 2 points clés fils. Par conséquent, il n'est pas nécessaire de construire une subdivision qui n'apporte pas d'information supplémentaire.

Un niveau de profondeur maximum trop faible crée de trop grands voisinage, regroupant de trop nombreux points clés. D'autre part, un arbre trop profond crée des grappes avec trop peu de voisins voire uniquement le point clé étudié. La cohérence de la correction géométrique ne serait alors pas pertinente.

Par la suite, chaque image comptant en moyenne 2 000 points clés dans les bases de données utilisées, nous considérons un niveau maximum de 5 pour la construction des arbres de recherche. En effet, dans ce cas l'image est divisée en un nombre maximal de $2^5 \times 2^5$ rectangles, c'est-à-dire 1 024 sous-zones.

VI.4. Influence de la correction géométrique des points clés

Etudions à présent l'influence de la correction géométrique sur les descripteurs retenus.

VI.4.1. Diminution du nombre de descripteurs retenus

Le **Tableau 9** présente le nombre de descripteurs catégorisé dans la classe *bâtiment*, avec et sans correction géométrique. Nous pouvons observer que la méthode de correction géométrique réduit le nombre de descripteurs conservés après filtrage notamment avec la classification retenue précédemment, *i.e.* en faisant intervenir différents classifieurs adaptés aux différentes catégories de bâtiments de la base de données et en sélectionnant le plus approprié selon le critère de probabilité maximale.

Cela s'explique par le fait que le nombre de descripteurs de la classe *non-bâtiment* est plus important dans ce cas-là que le nombre de descripteurs de la classe *bâtiment*. De ce fait, l'influence de la première sur la seconde est plus importante lors de la correction par voisinage.

		Sans correction	Avec correction
Unique classifieur SVM global	**Min**	612	553
	Max	3025	3462
	Moyenne	1403	1449
	Écart-type	625	752
Classifieurs SVM adaptés choisis selon le nombre	**Min**	280	137
	Max	2374	2543
	Moyenne	690	537
	Écart-type	561	652
Classifieurs SVM adaptés choisis selon la probabilité	**Min**	192	38
	Max	2374	2474
	Moyenne	612	422
	Écart-type	586	666

Tableau 9. Nombre moyen par image de descripteurs catégorisés dans la classe *bâtiment* par différentes classifications SVM.

VI.4.2. Diminution du nombre de faux positifs

D'autre part, le nombre de faux positifs diminue aussi en moyenne. Cela montre l'intérêt de la correction géométrique de façon à rejeter de façon optimale les descripteurs ne représentant pas les différents bâtiments à rechercher.

Les statistiques présentées dans les **Tableaux 10** et **11** ont été recensées sur les images de la vérité terrain de la base de données Paris6k. Les points classifiés en tant que *bâtiment* mais ne faisant en réalité pas partie du bâtiment recherchés sont comptabilisés comme faux positifs. Les ratios présentés dans le **Tableau 11** présente la diminution relative du nombre de faux positifs après correction géométrique, ce qui se traduit par une cohérence de la répartition géométrique des points clés selon leur appartenance aux classes *bâtiment* et *non-bâtiment*.

		Sans correction	Avec correction
Unique classifieur SVM global	**Min**	200	171
	Max	1198	1050
	Moyenne	465	384
	Écart-type	255	235
Classifieurs SVM adaptés choisis selon le nombre	**Min**	43	4
	Max	489	217
	Moyenne	165	55
	Écart-type	114	59
Classifieurs SVM adaptés choisis selon la probabilité	**Min**	8	3
	Max	287	129
	Moyenne	91	30
	Écart-type	80	42

Tableau 10. Nombre moyen par image de faux positifs prédits par différents classifieurs SVM.

		Sans correction	Avec correction
Unique classifieur SVM global	**Moyenne**	0,373	0,315
	Écart-type	0,180	0,182
Classifieurs SVM adaptés choisis selon le nombre	**Moyenne**	0,323	0,177
	Écart-type	0,208	0,173
Classifieurs SVM adaptés choisis selon la probabilité	**Moyenne**	0,210	0,164
	Écart-type	0,184	0,232

Tableau 11. Ratio moyen par image du nombre de faux positifs en fonction du nombre total de descripteurs prédits dans la classe *bâtiment* par différentes classifications SVM.

De manière générale, nous pouvons remarquer que le nombre de points clés retenus après filtrage est moins important. D'autre part, le nombre de faux positifs obtenus par la classification SVM en classes *bâtiment* et *non-bâtiment* est aussi moins important, permettant ainsi de rejeter plus de points ne faisant pas partie du bâtiment recherché sur l'image.

VI.5. Conclusion

Nous avons donc proposé une méthode permettant d'uniformiser la classification de chaque point clés suivant les classes *bâtiment* et *non-bâtiment* de façon à retrouver une cohérence

géométrique. Cette approche s'appuie sur la classification des voisins de chaque point de façon à prendre en compte cette information de classification apportée par les voisins.

Nous avons d'autre part défini une construction de recherche de voisinage efficace basée sur l'utilisation de *quadtree*.

Grâce à cette correction de voisinage géométrique pour chaque descripteur visuel classé, l'attribution de tous les points clés aux classes *bâtiment* et *non-bâtiment* est maintenant obtenue par une vérification et donc une cohérence géométrique.

Quelques exemples de résultats obtenus avec et sans correction géométrique sont présentés **Figure 89** à **Figure 94**.

Nous pouvons observer que la correction géométrique permet d'obtenir de nuages de points bien plus cohérents du point de vue de leur localisation spatiale dans l'image.

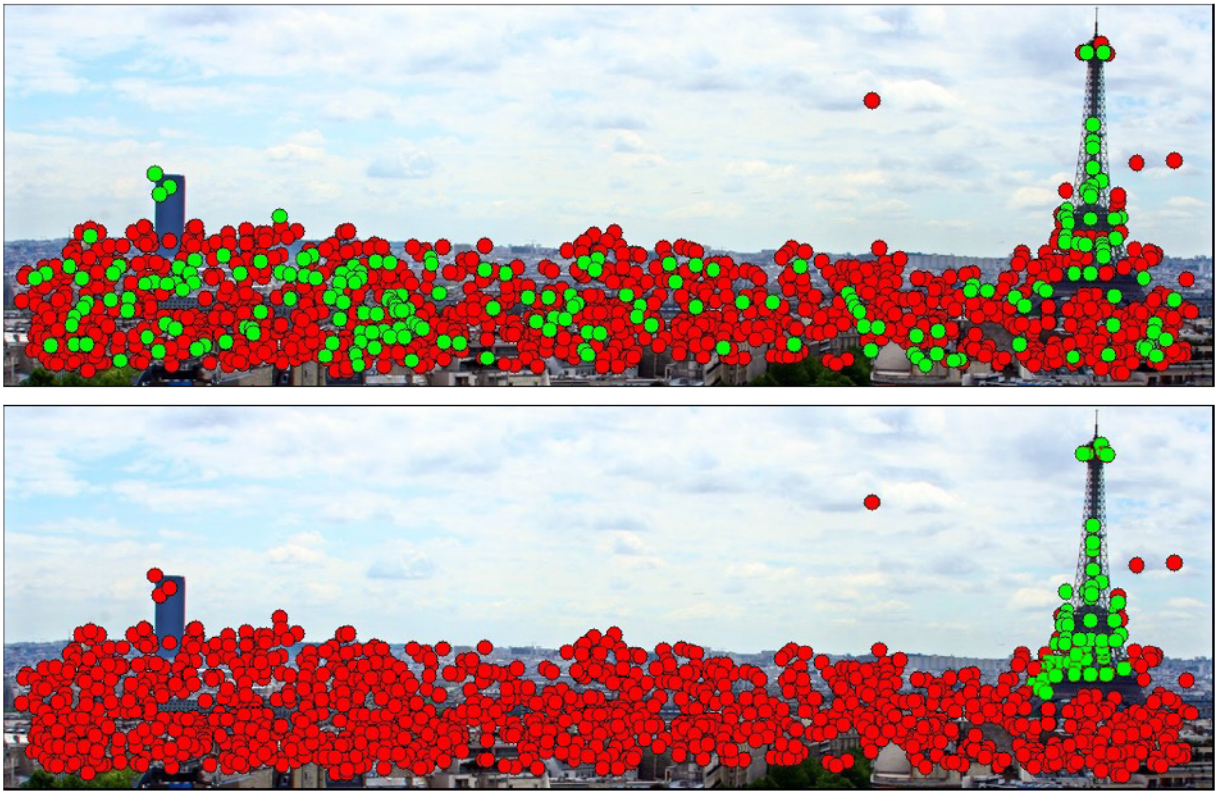


Figure 89. Exemple 1 de classification des points clés d'une image de la catégorie de bâtiment Tour Eiffel de la base de données Paris6k sans et avec correction géométrique.

Figure 90. Exemple 2 de classification des points clés d'une image de la catégorie de bâtiment Tour Eiffel de la base de données Paris6k sans et avec correction géométrique.

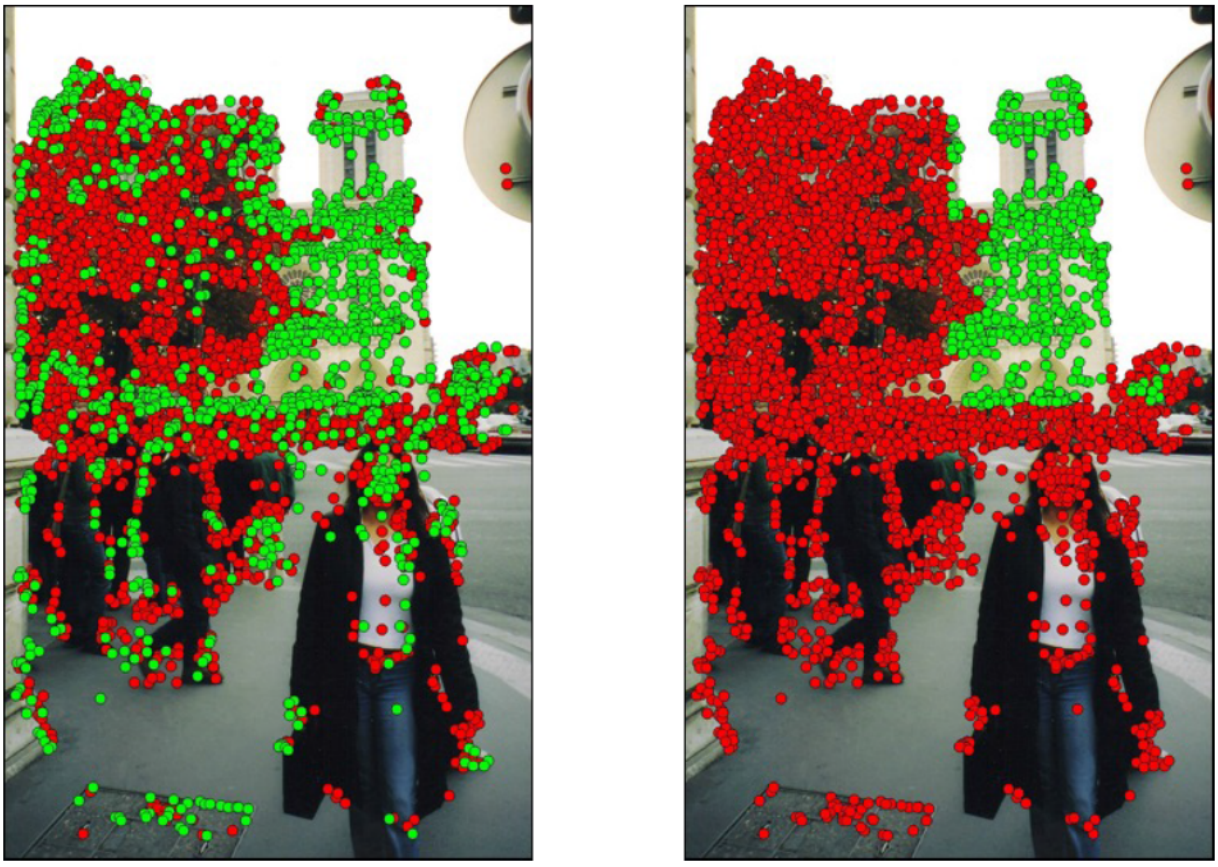


Figure 91. Exemple 1 de classification des points clés d'une image de la catégorie de bâtiment Notre Dame de la base de données Paris6k sans et avec correction géométrique.

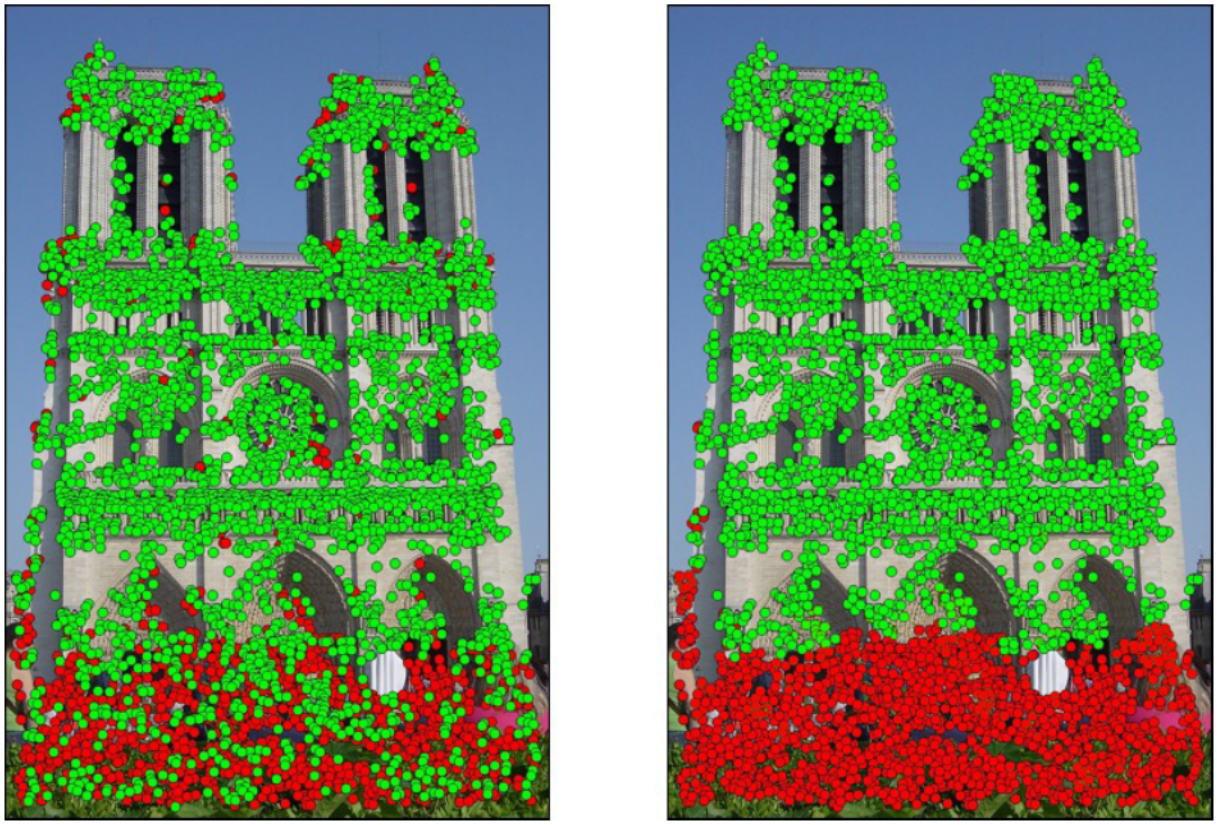


Figure 92. Exemple 2 de classification des points clés d'une image de la catégorie de bâtiment Notre Dame de la base de données Paris6k sans et avec correction géométrique.

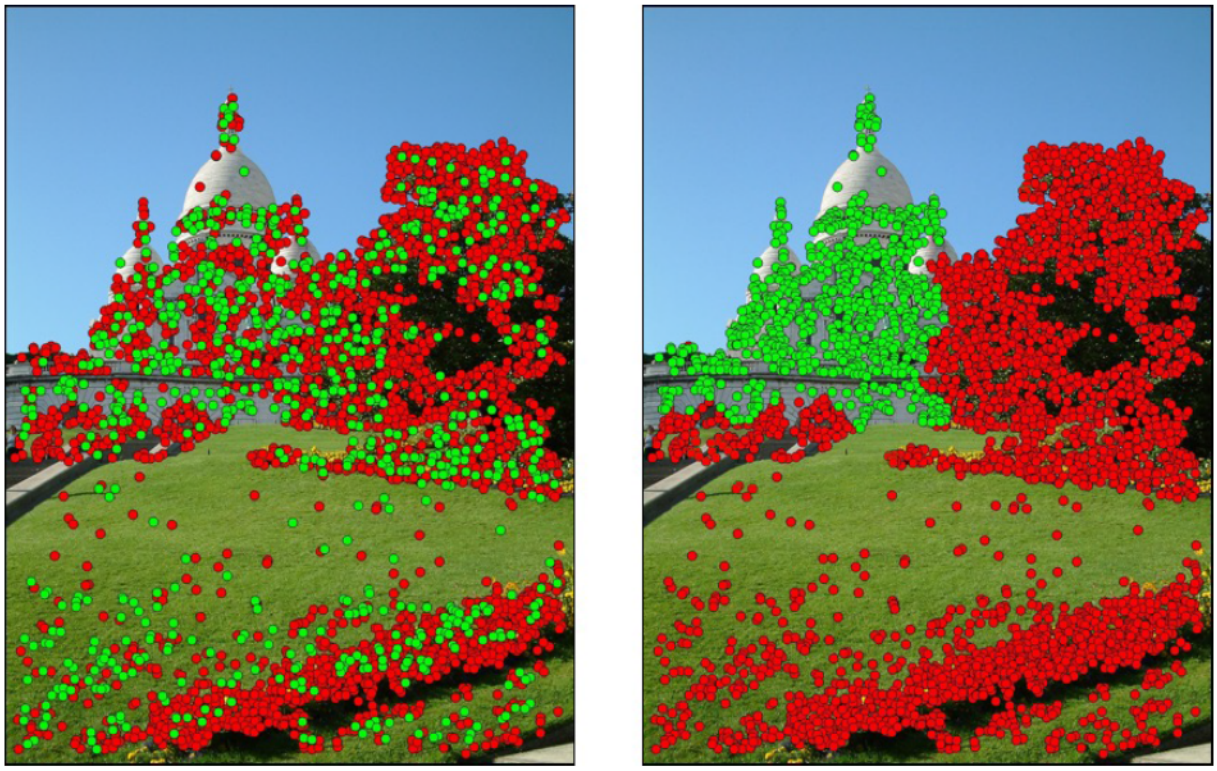


Figure 93. Exemple 1 de classification des points clés d'une image de la catégorie de bâtiment Sacré Cœur de la base de données Paris6k sans et avec correction géométrique.

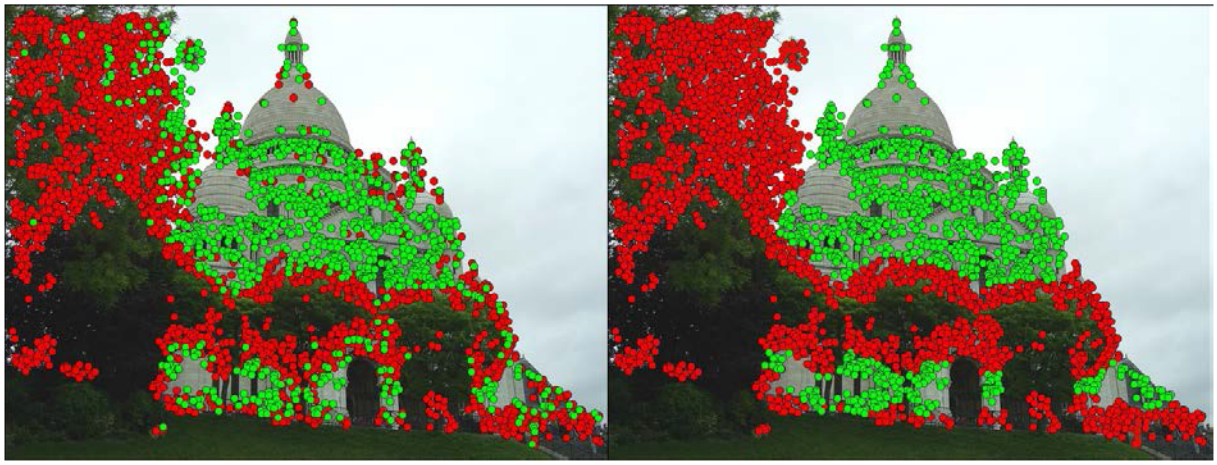


Figure 94. Exemple 2 de classification des points clés d'une image de la catégorie de bâtiment Sacré Cœur de la base de données Paris6k sans et avec correction géométrique.

Enfin, nous avons étudié le nombre de descripteurs retenu après la correction géométrique selon les différentes approches de classification proposées. Les résultats sont résumés dans le **Tableau 12**. Par comparaison, les mêmes résultats sont donnés **Tableau 13**.

	Paris6k	**Oxford5k**
État de l'art (sans filtrage)	2 506,79	2 669,70
Classifieur global unique	1 132,58	1 991,82
Classifieur choisi selon le nombre maximum de points clés retenus	203,79	106,29
Classifieur choisi selon la probabilité optimale de prédiction	121,08	56,83

Tableau 12. Nombre moyen de descripteurs retenus après la correction géométrique.

	Paris6k	**Oxford5k**
État de l'art (sans filtrage)	2 506,79	2 669,70
Classifieur global unique	1 165,29	1 820,20
Classifieur choisi selon le nombre maximum de points clés retenus	444,67	320,03
Classifieur choisi selon la probabilité optimale de prédiction	373,49	282,32

Tableau 13. Nombre moyen de descripteurs retenus sans correction géométrique.

Même si l'approche de correction géométrique permet de délimiter d'une manière plus fine et précise les bâtiments d'intérêt, il souffre d'une limitation liée au nombre trop réduit de points d'intérêt retenus. Cela est pénalisant pour la mise en œuvre des techniques de recherche par similarité, car il n'est pas possible de globaliser de manière fiable l'information à partir d'un nombre aussi réduit de points d'intérêt et descripteurs associés.

VII. RESULTATS EXPERIMENTAUX

Résumé. Ce chapitre présente les résultats expérimentaux obtenus avec les modèles BOW et les descripteurs VLAD dans le cadre de la recherche d'images par similarité. Les méthodes de classification de descripteurs locaux présentés précédemment ont pour but d'améliorer les performances de recherche d'image. Nous nous proposons d'évaluer ici ces améliorations *via* différentes mesures définies dans une première section. Les paramètres optimaux du modèle BOW et des différents modèles SVM entraînés sont choisis de façon empirique et présentés dans les sections suivantes. Nous montrons ainsi les améliorations apportées par les différentes approches de classification et de sélection des données locales proposées ainsi que l'influence de la correction géométrique proposée au **Chapitre VI**.

Mots clés : classification SVM, recherche d'image, image requête, représentation globale d'image, modèle BOW, descripteur VLAD, rappel et précision.

Les expérimentations ont été réalisées dans le cadre de recherches d'images sur les bases de données décrites au **Chapitre III**, qui recensent des monuments touristiques des villes de Paris [PCIS08] et d'Oxford[PCIS07]. Les vérités terrain fournie pour chacune de ces deux bases de données permettent de mettre en place une évaluation objective, en utilisant les critères d'évaluation largement répandus dans l'état de l'art, qui sont rappelés dans la section suivante.

VII.1. Mesures d'évaluation des résultats de recherche d'images

Parmi les métriques d'évaluation les plus populaires dans le domaine de l'indexation, citons les scores de rappel et de précision.

Le *rappel* correspond au rapport du nombre d'images correctement détectées renvoyées par le système en fonction du nombre total d'images qui aurait dû être retournées d'après la vérité terrain, comme décrit dans l'équation (27) :

$$rappel = \frac{|Correctes\ renvoyées|}{|Correctes\ selon\ la\ vérité\ terrain|} \tag{27}$$

La *précision* correspond au nombre d'images correctement détectées renvoyées par le système en fonction du nombre total d'images effectivement retournées en résultat comme défini dans l'équation (28).

$$précision = \frac{|Correctes\ renvoyées|}{|Total\ renvoyées|} \tag{28}$$

Pour combiner les deux mesures de rappel et de précision dans un score unique, il est habituel d'évaluer leur moyenne harmonique, nommée *F-score* et définie comme décrit dans l'équation (29).

$$F = 2 \times \frac{pécision \times rappel}{précision + rappel} \tag{29}$$

Enfin, nous retenons également la mesure de précisions moyenne (en anglais *Mean Average Precision* – MAP) définie dans l'équation (30) et évaluée par l'algorithme proposée par [PCIS07]. Cette mesure MAP est basée sur la mesure de précision moyenne (en anglais *Average Precision*, AP) définie dans l'équation (31). Dans ces équations, Q est le nombre d'images requêtes, $P(k)$ est la

mesure de précision pour la liste d'images retournées jusqu'au rang k et $\delta(k)$ est une fonction d'indication valant 1 si l'image retournée au rang k est correcte et 0 sinon.

$$MAP = \frac{\sum_{k=1}^{Q} AP(k)}{Q} \tag{30}$$

$$AP(n) = \frac{\sum_{k=1}^{n} P(k) \times \delta(k)}{|Correctes|} \tag{31}$$

La méthode de sélection du classifieur SVM adapté présentée au **Chapitre V** permet non seulement de filtrer de façon plus pertinente les points clés retenus pour construire notre vocabulaire, mais d'autre part de catégoriser implicitement l'image requête dans l'une des catégories de bâtiments définies dans la vérité de terrain, en s'appuyant sur le classifieur le plus adapté identifié. En effet, pour une image donnée nous évaluons la classification des points clés en *bâtiment* et *non-bâtiment* selon les différents classifieurs SVM adaptés aux différentes catégories. Selon nos deux critères de sélection définis au **Chapitre V**, l'objectif est de choisir de façon automatique le classifieur SVM ayant été entraîné avec les images de la catégorie correspondant effectivement à l'image donnée. Par exemple, pour une image de la base de données Paris6k représentant la Tour Eiffel, le classifieur adapté est celui entraîné avec les images de la Tour Eiffel, puisque ce dernier filtre de façon plus précise et pertinente les points clés du monument. Par conséquent, nous obtenons ainsi l'information du classifieur SVM automatiquement sélectionné le plus adapté à l'image requête, ce qui est équivalent à reconnaître la catégorie de bâtiments des images exploitées pour l'entraînement de ce classifieur. Cela nous permet donc d'en déduire la catégorie de bâtiment à laquelle appartient notre image requête.

Le reste de ce chapitre est structuré comme suit. En premier lieu, nous étudions l'influence des différents paramètres qui interviennent dans le processus de recherche au niveau du modèle BOW (**Paragraphe VII.2**). Le **Paragraphe VII.3** présente les résultats de recherche d'images de l'état de l'art avec les modèles usuels de BOW et VLAD. Nous introduisons ensuite les résultats de cette même recherche d'images selon la méthode de filtrage globale, faisant intervenir un simple modèle de classification SVM entraîné de façon globale, sans distinction des catégories de bâtiments (**Paragraphe VII.4**). Par la suite nous exposons ces mêmes résultats obtenus avec un classifieur adapté à chaque catégorie de bâtiment selon les deux critères de choix définis au **Chapitre V** (**Paragraphe VII.5**). De plus, les différentes méthodes de filtrage sont accompagnées des résultats représentant l'influence de la correction géométrique présentée au **Chapitre VI**, en termes de rappel et de précision dans le dernier paragraphe dressant un bilan de l'ensemble des résultats obtenus.

Intéressons-nous à présent aux paramètres caractérisant le modèle BOW.

VII.2. Paramètres du modèle de BOW

Le modèle BOW s'appuie sur la construction d'un vocabulaire de *mots visuels*, construit à partir des données d'apprentissage par des procédures plus ou moins classiques de *clustering*. Dans ce cadre, le paramètre primordial à prendre en compte dans la construction du modèle concerne le nombre de mots visuels de ce vocabulaire.

Pour déterminer la taille du vocabulaire, nous avons considéré une expérimentation de recherche d'images par simple modèle BOW, conduite sur la base de données Paris6k. Dans ce cadre, nous avons étudié l'évolution du score de précision moyenne MAP en fonction des différentes tailles du vocabulaire.

Les vocabulaires considérés sont de 100, 1 000, 2 000, 2 100, 2 200, 2 300, 2 400, 2 500, 2 600, 2 700, 2 800, 2 900, 3 000, 4 000, 5 000, 7 500, 10 000, 20 000 et 250 000 mots.

Les résultats obtenus sont présentés **Figure 95** (scores MAP) et **Figure 96** (évolution en % des scores MAP).

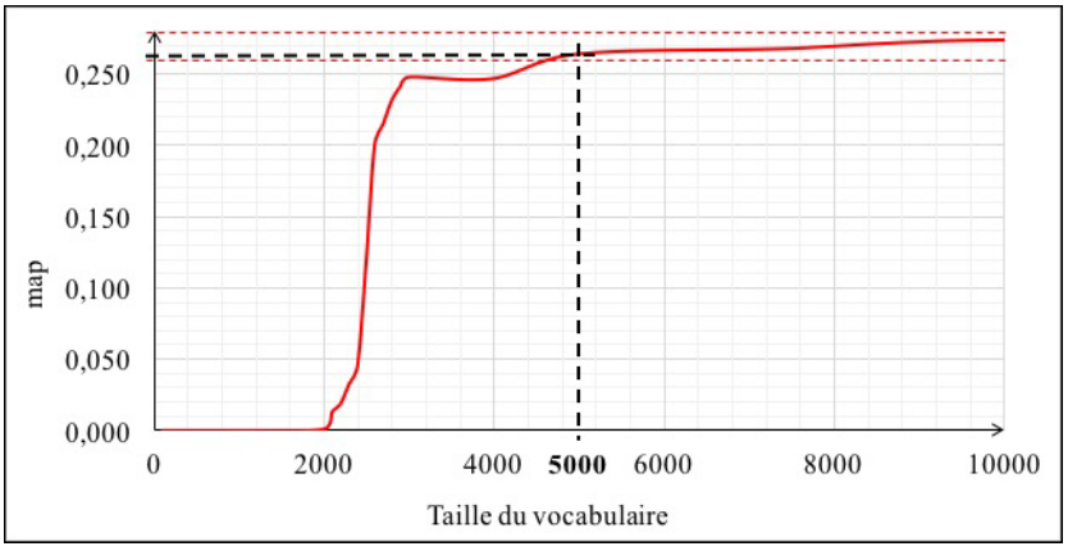


Figure 95. Évolution de la précision moyenne (score MAP) de recherche d'image avec le modèle de BOW en fonction de la taille de vocabulaire définie.

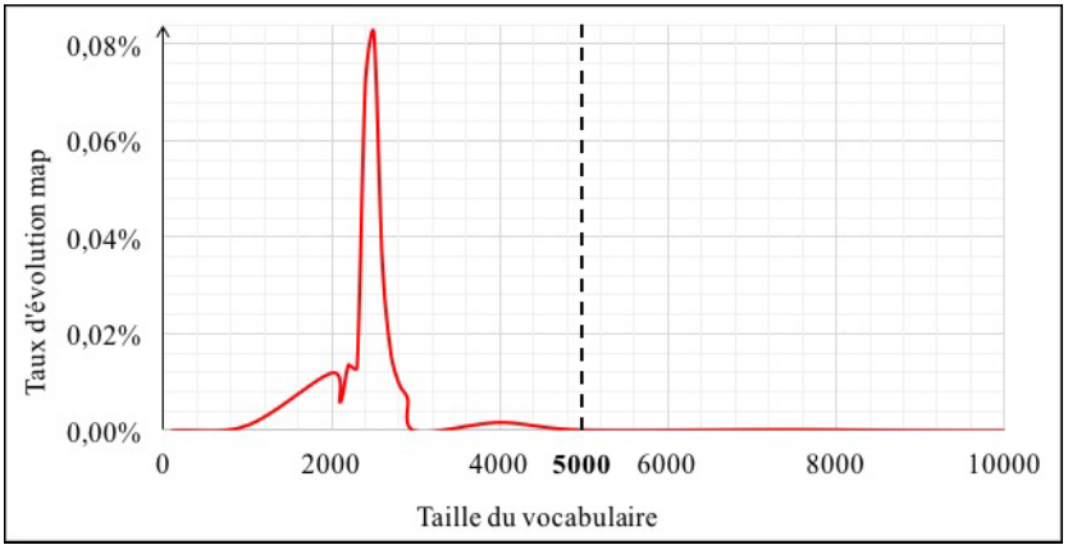


Figure 96. Taux d'évolution de la précision moyenne (score MAP) de recherche d'image avec le modèle BOW en fonction de la taille de vocabulaire définie.

D'après le premier graphique, nous pouvons d'ors et déjà remarquer qu'un vocabulaire contenant moins de 2 000 mots ne retourne quasiment aucun résultat correctement identifié. En effet, pour une taille de vocabulaire faible, le nombre de mots n'est pas suffisant pour représenter les plus de 16 millions points d'intérêt extraits à partir des 6 412 images de la base de données Paris6k. Dans ce cas, les images retrouvées par recherche de bâtiment pour les images requêtes de la vérité terrain ne fournissent pas de résultats précis et pour chaque catégorie, très peu d'images font effectivement partie des images correctement détectées.

D'autre part, à partir d'un vocabulaire de 3 000 mots, les résultats de précision moyenne atteignent un palier et son taux évolution ne varie que d'une manière marginale.

Enfin, au-delà d'un vocabulaire de 6 000 mots, le temps nécessaire à la création du vocabulaire devient trop important pour les machines utilisées lors des expérimentations pour une évolution des résultats de précision tout à fait négligeable.

Notons que de résultats tout à fait équivalents ont été également obtenus sur la base de données Oxford5k.

Pour toutes ces raisons, dans la suite de nos développements, nous avons retenu une taille de vocabulaire de 5 000 mots visuels, pour l'ensemble des modèles BOW et quelle que soit la base de données considérée.

Exposons à présent les résultats de recherche d'images obtenus. En premier lieu, intéressons-nous aux modèles classiques par représentations aussi bien BOW et VLAD, qui servent de *baseline* pour nos comparaisons.

VII.3. Résultats de la recherche d'images dans l'état de l'art

La grande majorité des résultats rapportés dans l'état de l'art, exposent uniquement la précision moyenne (le score MAP). De façon à pouvoir comparer objectivement les résultats des méthodes précédemment présentées, les résultats de l'état de l'art ont été réévalués par notre propre implantation, pour garantir l'utilisation des mêmes paramètres, notamment en termes de taille de vocabulaire utilisé. Nous avons retenu pour évaluation l'ensemble des mesures présentées au **Paragraphe VII.1** : précision, rappel, *F-score* et score *MAP*.

Les résultats de référence obtenus pour la recherche d'images avec les modèles de BOW (**Tableau 14** et **Tableau 15**) et VLAD (**Tableau 16**) sont présentés sur les deux bases de données considérées, Paris6k et Oxford5k.

Comme tendance globale, nous pouvons d'ores et déjà remarquer que le score MAP global de la recherche d'images avec le modèle VLAD est supérieur au score obtenu avec la représentation BOW. Ainsi, nous observons une amélioration du score MAP de 15,4% entre les résultats du BOW par rapport aux résultats du VLAD pour la base de données Paris6k et de 19,6% pour la base de données Oxford5k.

Cela montre sans surprise la supériorité, en termes de pouvoir de discrimination, de la représentation VLAD par rapport au modèle BOW.

Quelques exemples des premières images retournées sont présentées pour le cas d'utilisation de l'état de l'art avec la même image requête sont présentés **Figure 97** et **Figure 98**. Dans ces exemples, l'image requête est la première image retournée.

Figure 97. Exemples de résultat de recherche d'images par le modèle de BOW pour la catégorie de bâtiment représentant la Tour Eiffel sans filtrage par classification.

Figure 98. Exemples de résultat de recherche d'images par le modèle de BOW pour la catégorie de bâtiment représentant les Invalides sans filtrage par classification.

Notons cependant que les images retournées en premières positions sont peu satisfaisantes : une seule image pour la requête correspondant à la Tour Eiffel et aucune pour celle représentant le Dôme des Invalides, dans les premiers cinq résultats retournés.

Ces mêmes exemples de requêtes seront repris dans les sections suivantes pour illustrer l'apport des techniques de filtrage de points d'intérêt proposées.

Catégorie	Rappel	Précision	F-mesure	MAP
La Défense	0,191	0,198	0,195	0,137
Tour Eiffel	0,276	0,338	0,303	0,205
Invalides	0,323	0,328	0,325	0,210
Louvre	0,476	0,606	0,531	0,412
Moulin Rouge	0,440	0,468	0,454	0,393
Musée d'Orsay	0,372	0,382	0,377	0,301
Notre Dame	0,681	0,760	0,718	0,656
Panthéon	0,665	0,789	0,721	0,642
Pompidou	0,835	0,951	0,889	0,831
Sacré Cœur	0,638	0,770	0,697	0,598
Arc de Triomphe	0,364	0,390	0,377	0,325
Moyenne	**0,478**	**0,544**	**0,508**	**0,428**

Tableau 14. Résultats de la recherche d'images dans la base de données Paris6k avec la méthode de BOW de l'état de l'art.

Catégorie	Rappel	Précision	F-mesure	MAP
All Souls	0,544	0,553	0,548	0,495
Ashmolean	0,272	0,272	0,272	0,211
Balliol	0,517	0,545	0,530	0,411
Bodleian	0,517	0,558	0,536	0,499
Christ Church	0,490	0,506	0,498	0,442
Cornmarket	0,400	0,422	0,410	0,374
Hertford	0,463	0,465	0,464	0,404
Keble	0,714	0,714	0,714	0,705
Magdalen	0,074	0,075	0,074	0,044
Pitt Rivers	0,667	0,793	0,724	0,659
Radcliffe Camera	0,593	0,629	0,610	0,511
Moyenne	**0,477**	**0,503**	**0,489**	**0,432**

Tableau 15. Résultats de la recherche d'images dans la base de données Oxford5k avec la méthode de BOW de l'état de l'art.

	Paris6k	Oxford5k
Score MAP	0,500	0,524

Tableau 16. Score MAP de la recherche d'images dans les bases de données Paris6k et Oxford5k avec le modèle de descripteur VLAD de l'état de l'art.

Selon le même protocole d'évaluation, analysons à présent l'apport des méthodes de filtrage proposées, en termes d'amélioration des performances des requêtes, en commençant par le filtrage global (*cf.* **Chapitre IV**), avec un unique classifieur SVM.

Dans l'état de l'art, tous les descripteurs locaux SIFT sont conservés pour construire ce vocabulaire. En utilisant nos deux méthodes de filtrage proposées basées sur une classification SVM des points d'intérêt, nous ne prenons en considération que les points attribués à la classe *bâtiment* pour chaque image filtrée pour rendre le vocabulaire plus adapté à la recherche effective des monuments spécifiques. Cependant, le filtrage des points clés pour certaines images ne retient qu'une dizaine de points clés. Afin d'éviter les erreurs de reconnaissance dues à un manque de données, un nouvel ensemble d'images est construit à partir de la base de données initiales en retirant les images présentant moins de 10 descripteurs après filtrage. Nous désignons par la suite ce sous-ensemble de la base initiale comme *base de données réduite*. Les résultats obtenus avec les différentes méthodes de filtrage de point d'intérêt seront donc rapportés sur cette base réduite. Le nombre d'images restant dépend donc du mode de filtrage choisi. Le **Tableau 17** présente le nombre d'images dans les différentes bases de données dites *réduites* suivant les différents critères de classifications pour les bases de données Paris6k et Oxford5k.

	Paris6k	Oxford5k
État de l'art (sans filtrage)	6 412	5 063
Classifieur global unique	6 381	5 057
Classifieur choisi selon le nombre maximum de points clés retenus	6 363	5 033
Classifieur choisi selon la probabilité optimale de prédiction	5 735	4 843

Tableau 17. Nombre d'images dans les bases de données dites *réduites* selon les différentes stratégies de classifications SVM pour les bases de données Paris6k et Oxford5k.

VII.4. Résultats du filtrage des descripteurs locaux avec un unique classifieur global

Ici, les descripteurs locaux SIFT conservés pour la construction de notre vocabulaire sont filtrés via un classifieur SVM entraîné globalement pour l'ensemble des catégories d'images de la base de données.

Pour construire le vocabulaire, uniquement les descripteurs locaux classifiés dans la classe *bâtiment* sont considérés.

Ici, les scores de rappel, les scores de précision, les moyennes harmoniques *F-mesure* et les précisions *MAP* sont présentés dans les **Tableaux 18** et **19** pour les bases de données Paris6k et Oxford5k respectivement. Globalement, nous pouvons observer une amélioration encourageante des résultats de recherche d'image grâce à la réduction du nombre de mots clés pris en compte pour la construction du vocabulaire, qui permet de conserver seulement les mots clés les plus adaptés pour décrire les objets d'intérêt recherchés. Ainsi, les gains obtenus en termes de score MAP sont de 0,10% pour la base Paris6k et de 0,08% pour celle d'Oxford5k, par rapport à la méthode BOW de référence.

Des exemples de résultats de recherche d'images sont présentées **Figure 99** et **Figure 100** par ordre de similarité décroissante avec la requête. Nous observons ici une nette amélioration des résultats obtenus par rapport à la méthode de base.

Figure 99. Exemples de résultat de recherche d'images par le modèle de BOW pour la catégorie de bâtiment représentant la Tour Eiffel sans filtrage par classification et *via* un filtrage avec un modèle SVM global.

Figure 100. Exemples de résultat de recherche d'images par le modèle de BOW pour la catégorie de bâtiment représentant les Invalides sans filtrage par classification et *via* un filtrage avec un modèle SVM global.

Catégorie	Rappel	Précision	F-mesure	MAP
La Défense	0,202	0,211	0,206	0,157
Tour Eiffel	0,331	0,416	0,368	0,262
Invalides	0,553	0,568	0,560	0,492
Louvre	0,426	0,575	0,489	0,347
Moulin Rouge	0,391	0,403	0,397	0,312
Musée d'Orsay	0,269	0,270	0,270	0,204
Notre Dame	0,761	0,856	0,805	0,749
Panthéon	0,748	0,868	0,803	0,727
Pompidou	0,847	0,955	0,898	0,841
Sacré Cœur	0,770	0,954	0,852	0,767
Arc de Triomphe	0,410	0,442	0,425	0,370
Moyenne	**0,519**	**0,592**	**0,552**	**0,475**

Tableau 18. Résultats de la recherche d'images dans la base de données Paris6k avec la méthode de BOW après sélection des points clés par classifieur SVM global.

Catégorie	Rappel	Précision	F-mesure	MAP
All Souls	0,585	0,591	0,588	0,534
Ashmolean	0,256	0,258	0,257	0,177
Balliol	0,533	0,552	0,542	0,455
Bodleian	0,500	0,549	0,523	0,464
Christ Church	0,521	0,543	0,532	0,467
Cornmarket	0,444	0,483	0,463	0,440
Hertford	0,522	0,526	0,524	0,453
Keble	0,657	0,657	0,657	0,652
Magdalen	0,093	0,093	0,093	0,056
Pitt Rivers	0,800	0,960	0,873	0,800
Radcliffe Camera	0,719	0,767	0,742	0,688
Moyenne	**0,512**	**0,544**	**0,527**	**0,472**

Tableau 19. Résultats de la recherche d'images dans la base de données Oxford5k avec la méthode de BOW après sélection des points clés grâce à notre classifieur SVM global.

Ces résultats montrent que la sélection appropriée des points d'intérêt utilisés pour la représentation d'image a un impact positif sur les performances des requêtes. Sur la même lignée, étudions maintenant l'impact du modèle de filtrage plus élaboré, à base de classifieurs SVM multiples.

VII.5. Résultats du filtrage des descripteurs locaux avec différents modèles SVM adaptés à chaque catégorie

Dans cette section, nous présentons les résultats obtenus pour la recherche d'images avec un vocabulaire construit selon notre seconde méthode de sélection des points clés. Ici, les descripteurs locaux SIFT conservés pour la construction de notre vocabulaire sont filtrés selon différents classifieurs SVM entraînés de façon adaptée à chacune des catégories d'image de la base de données, *cf.* méthode présentée au **Chapitre** V.

Nous présentons les résultats obtenus suivant les deux approches de sélection du classifieur adapté présentées dans le **Chapitre V**. Le premier critère de sélection fait intervenir le nombre de descripteurs locaux conservés par les différents classifieurs entraînés tandis que le second critère prend en compte le score de confiance de la prédiction fourni par les différents classifieurs. Les deux approches de représentation, BOW et VLAD sont ici considérées.

VII.5.1. Recherche d'images par modèle de BOW

VII.5.1.1. Choix du classifieur SVM suivant le critère du nombre maximum de points clés retenus

Dans ce paragraphe, le classifieur choisi pour le filtrage parmi les différents entraînés est celui retenant le plus grand nombre de points clés dans la classe *bâtiment*. Les résultats obtenus sont présentés dans les **Tableau 20** (pour la base Paris6k) et **Tableau 21** (pour la base Oxford5k) pour les scores de rappel, les scores de précision, les moyennes harmoniques *F-mesure* et les précisions *MAP*.

Les scores obtenus sont quasi-équivalents à ceux issus de la sélection des descripteurs avec un unique classifieur entraîné globalement, bien que légèrement supérieurs (de 0,05% pour la base Paris6k et 0,16% pour la base Oxford5k). Ce phénomène peut être expliqué par le fait qu'en privilégiant le classifieur qui fournit un nombre maximum de points d'intérêts le nombre de descripteurs retenus est globalement similaire à celui correspondant un unique classifieur global.

Des exemples d'images retournées lors de la recherche d'images par requête sont présentés **Figure 101** et **Figure 102**.

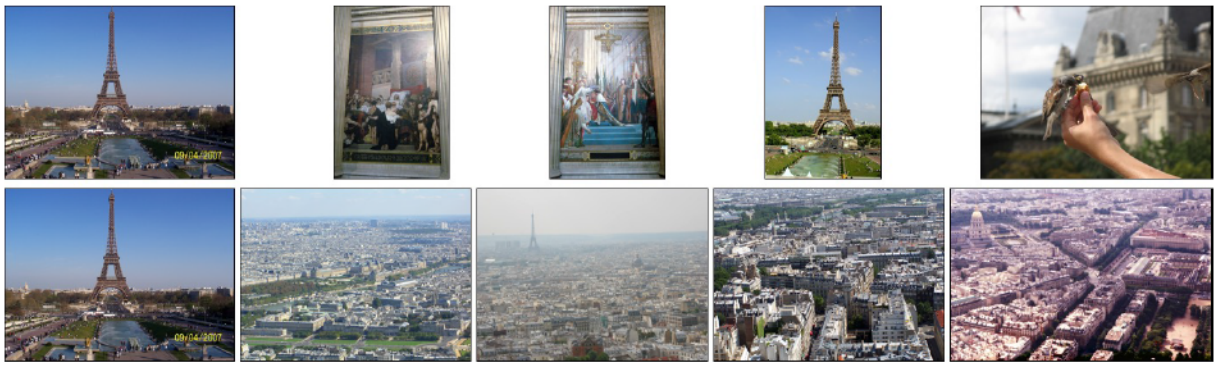


Figure 101. Exemples de résultat de recherche d'images par le modèle de BOW pour la catégorie de bâtiment représentant la Tour Eiffel sans filtrage par classification et *via* un filtrage avec différents modèles SVM adaptés au différentes catégories d'images et sélectionné selon le critère de maximum de points clés retenus.

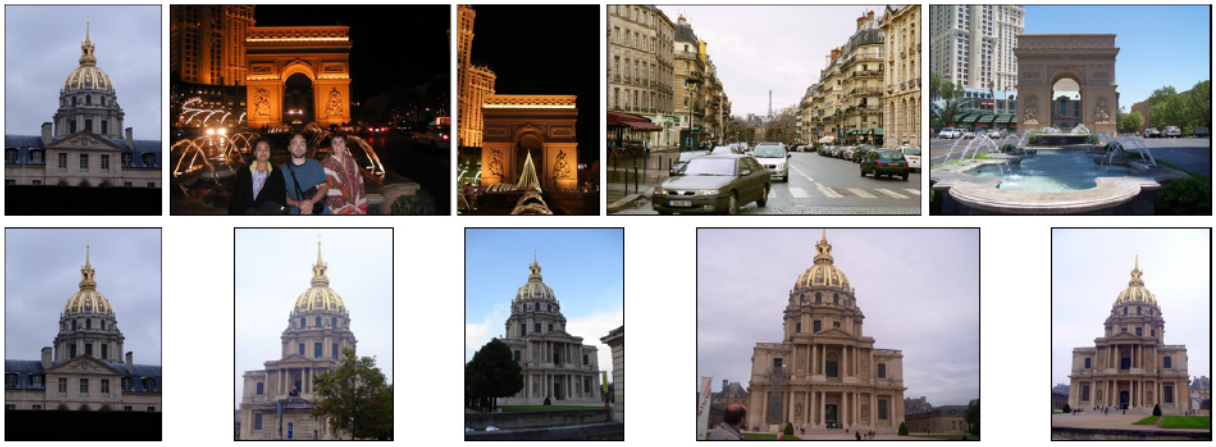


Figure 102. Exemples de résultat de recherche d'images par le modèle de BOW pour la catégorie de bâtiment représentant les Invalides sans filtrage par classification et *via* un filtrage avec différents modèles SVM adaptés au différentes catégories d'images et sélectionné selon le critère de maximum de points clés retenus.

Catégorie	Rappel	Précision	F-mesure	MAP
La Défense	0,108	0,112	0,110	0,079
Tour Eiffel	0,237	0,280	0,257	0,195
Invalides	0,482	0,492	0,487	0,473
Louvre	0,354	0,501	0,414	0,295
Moulin Rouge	0,219	0,219	0,219	0,163
Musée d'Orsay	0,306	0,316	0,310	0,262
Notre Dame	0,787	0,919	0,847	0,779
Panthéon	0,749	0,906	0,820	0,748
Pompidou	0,812	0,941	0,871	0,803
Sacré Cœur	0,768	0,981	0,861	0,767
Arc de Triomphe	0,411	0,429	0,420	0,373
Moyenne	**0,475**	**0,554**	**0,511**	**0,449**

Tableau 20. Résultats de la recherche d'images dans la base de données Paris6k avec la méthode de BOW après filtrage des points clés grâce à différents classifieurs SVM adaptés et sélection du classifieur par nombre de points clés retenus.

Catégorie	Rappel	Précision	F-mesure	MAP
All Souls	0,713	0,733	0,723	0,631
Ashmolean	0,344	0,344	0,344	0,286
Balliol	0,467	0,483	0,475	0,438
Bodleian	0,458	0,502	0,479	0,394
Christ Church	0,546	0,574	0,560	0,478
Cornmarket	0,578	0,633	0,604	0,574
Hertford	0,515	0,519	0,517	0,461
Keble	0,829	0,829	0,829	0,829
Magdalen	0,107	0,108	0,107	0,064
Pitt Rivers	0,800	0,933	0,861	0,783
Radcliffe Camera	0,786	0,840	0,812	0,747
Moyenne	**0,558**	**0,591**	**0,574**	**0,517**

Tableau 21. Résultats de la recherche d'images dans la base de données Oxford5k avec la méthode de BOW après filtrage des points clés grâce à différents classifieurs SVM adaptés et sélection du classifieur par nombre de points clés retenus.

Analysons à présent les résultats obtenus avec le même modèle BOW lorsque le deuxième critère de sélection de points d'intérêt, par vraisemblance globale de prédiction, est utilisé.

VII.5.1.2. Choix du classifieur SVM suivant le critère de la probabilité de prédiction optimale

Dans ce paragraphe, le classifieur choisi pour le filtrage parmi les différents entraînés est celui présentant la plus haute probabilité de prédiction de classe pour l'ensemble des points de la classe *bâtiment* d'une image donnée. Les scores de rappel, les scores de précision, les moyennes harmoniques *F-mesure* et les précision *MAP* sont présentées dans les **Tableau 22** et **Tableau 23**.

De manière générale, les scores obtenus dans ce cas sont bien supérieurs, pour les deux bases de test considérées, à ceux des méthodes de sélection précédentes. Cela démontre la pertinence de l'approche proposée et l'intérêt de disposer d'une méthode efficace de sélection de points clés lors de la phase de construction du vocabulaire.

D'autre part, cela nous montre que la classification fruste, binaire n'est pas toujours pertinente et qu'utiliser plutôt une mesure continue de cette information pour pondérer les résultats de la classification est plus pertinent.

Des exemples d'images retournées lors de la recherche d'images par requête sont présentés **Figure 103** et **Figure 104**.

Figure 103. Exemples de résultat de recherche d'images par le modèle de BOW pour la catégorie de bâtiment représentant la Tour Eiffel sans filtrage par classification et *via* un filtrage avec différents modèles SVM adaptés au différentes catégories d'images et sélectionné selon le critère de probabilité de prédiction optimal.

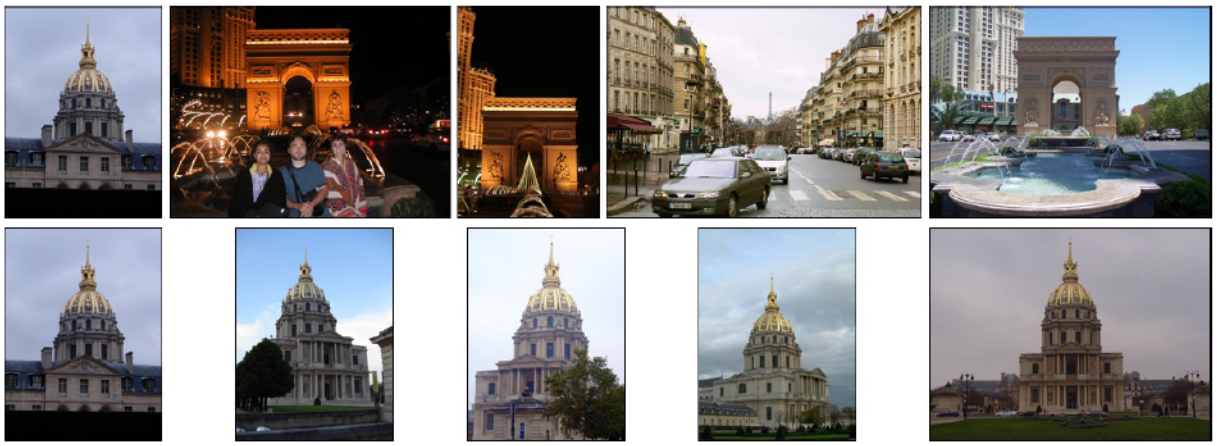


Figure 104. Exemples de résultat de recherche d'images par le modèle de BOW pour la catégorie de bâtiment représentant les Invalides sans filtrage par classification et *via* un filtrage avec différents modèles SVM adaptés au différentes catégories d'images et sélectionné selon le critère de probabilité de prédiction optimal.

Catégorie	Rappel	Précision	F-mesure	MAP
La Défense	0,248	0,252	0,250	0,188
Tour Eiffel	0,361	0,430	0,392	0,293
Invalides	0,573	0,585	0,579	0,559
Louvre	0,468	0,584	0,518	0,378
Moulin Rouge	0,321	0,329	0,325	0,279
Musée d'Orsay	0,319	0,324	0,322	0,275
Notre Dame	0,835	0,944	0,886	0,825
Panthéon	0,746	0,953	0,836	0,721
Pompidou	0,827	0,946	0,883	0,818
Sacré Cœur	0,789	0,987	0,877	0,789
Arc de Triomphe	0,431	0,459	0,445	0,384
Moyenne	**0,538**	**0,617**	**0,574**	**0,501**

Tableau 22. Résultats de la recherche d'images dans la base de données Paris6k avec la méthode de BOW après filtrage des points clés grâce à différents classifieurs SVM adaptés et sélection du classifieur par probabilité de prédiction.

Catégorie	Rappel	Précision	F-mesure	MAP
All Souls	0,677	0,701	0,689	0,610
Ashmolean	0,272	0,272	0,272	0,194
Balliol	0,467	0,474	0,470	0,442
Bodleian	0,433	0,470	0,451	0,379
Christ Church	0,538	0,565	0,551	0,473
Cornmarket	0,533	0,589	0,559	0,516
Hertford	0,519	0,523	0,520	0,458
Keble	0,829	0,829	0,829	0,824
Magdalen	0,126	0,126	0,126	0,090
Pitt Rivers	0,833	0,933	0,879	0,806
Radcliffe Camera	0,779	0,830	0,804	0,736
Moyenne	**0,546**	**0,574**	**0,559**	**0,502**

Tableau 23. Résultats de la recherche d'images dans la base de données Oxford5k avec la méthode de BOW après filtrage des points clés grâce à différents classifieurs SVM adaptés et sélection du classifieur par probabilité de prédiction.

En résumé, les résultats présentés montrent que la sélection des descripteurs locaux pour la construction du vocabulaire du modèle BOW permet d'améliorer les résultats de recherche d'image. Le filtre le plus performant prend en compte les classifieurs multiples, adaptés à chaque catégorie de la base de données, avec un critère de sélection fondé sur le score de vraisemblance.

Une synthèse des résultats obtenus avec la méthode BOW de façon indépendante des bases de données Paris6k et Oxford5k est présentée dans le **Tableau 24**.

	État de l'art	Unique classifieur SVM global	Classifieurs SVM adaptés choisis selon le nombre	Classifieurs SVM adaptés choisis selon la probabilité
Rappel	0,478	0,515	0,517	0,542
Précision	0,523	0,568	0,572	0,596
F-mesure	0,499	0,539	0,542	0,566
MAP	0,430	0,473	0,483	0,502

Tableau 24. Résumé des résultats de recherche d'image avec la méthode BOW indépendamment des bases de données utilisées.

VII.5.1.3. Temps de calcul

Nous présentons dans ce paragraphe les différents temps de calcul de la recherche d'images par requête sur les différentes bases de données avec le modèle BOW. Les **Figures 105** et **106** présentent les résultats obtenus en moyenne pour chaque catégorie de bâtiment. Chaque catégorie de bâtiment est représentée par cinq images requêtes définies dans la vérité terrain. Nous observons que le temps de calcul diminue fortement en considérant un filtrage des points d'intérêt avec différents modèles SVM adaptés. Ceci est à mettre en corrélation avec le nombre de points clés retenus après filtrage. En effet, les deux approches proposées avec un filtrage adapté aux différentes catégories de bâtiment de la base de données compte beaucoup moins de points d'intérêt que le modèle habituel de BOW et l'approche de filtrage avec un unique classifieur global.

En moyenne, pour la base de données Paris6k, le temps de recherche à partir d'une image requête avec le modèle BOW sans filtrage est de 5,71 secondes pour une image requête, de 5,33 secondes avec un unique classifieur filtrant les points d'intérêt et est réduit à 2,54 secondes et 2,03 secondes pour les filtrages avec les modèles SVM adaptés, respectivement selon le nombre maximum de points clés retenus et selon le score de confiance moyen optimal. Nous obtenons donc une amélioration du temps de calcul moyen par image de 64,4% avec l'approche proposée par rapport au modèle de l'état de l'art.

De même, pour la base de données Oxford5k, le temps de recherche à partir d'une image requête avec le modèle BOW sans filtrage est de 8,06 secondes pour une image requête, de 6,36 secondes avec un unique classifieur filtrant les points d'intérêt et est réduit à 1,72 secondes et 1,52 secondes pour les filtrages avec les modèles SVM adaptés, respectivement selon le nombre maximum de points clés retenus et selon le score de confiance moyen optimal. Nous obtenus donc une amélioration du temps de calcul moyen par image de 81,2% avec l'approche proposée par rapport au modèle de l'état de l'art.

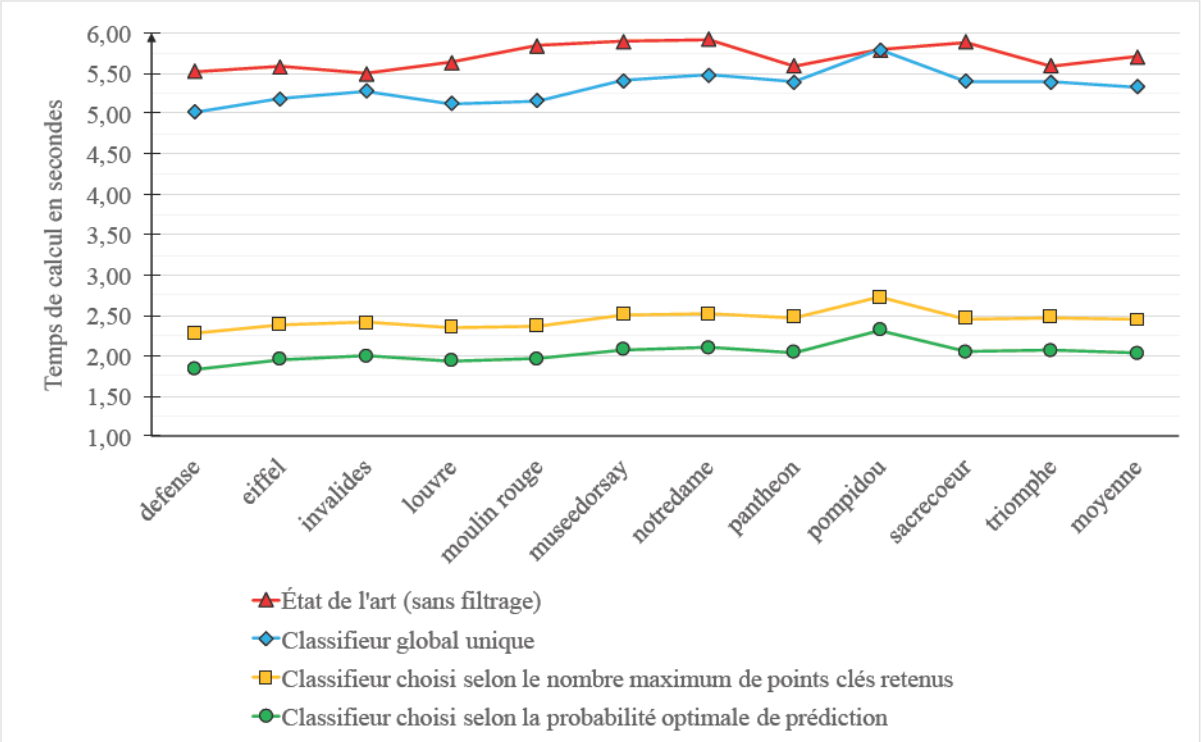


Figure 105. Temps de calcul moyen de recherches d'images par requête sur la base de données Paris6k avec les différentes méthodes de filtrage des points d'intérêt.

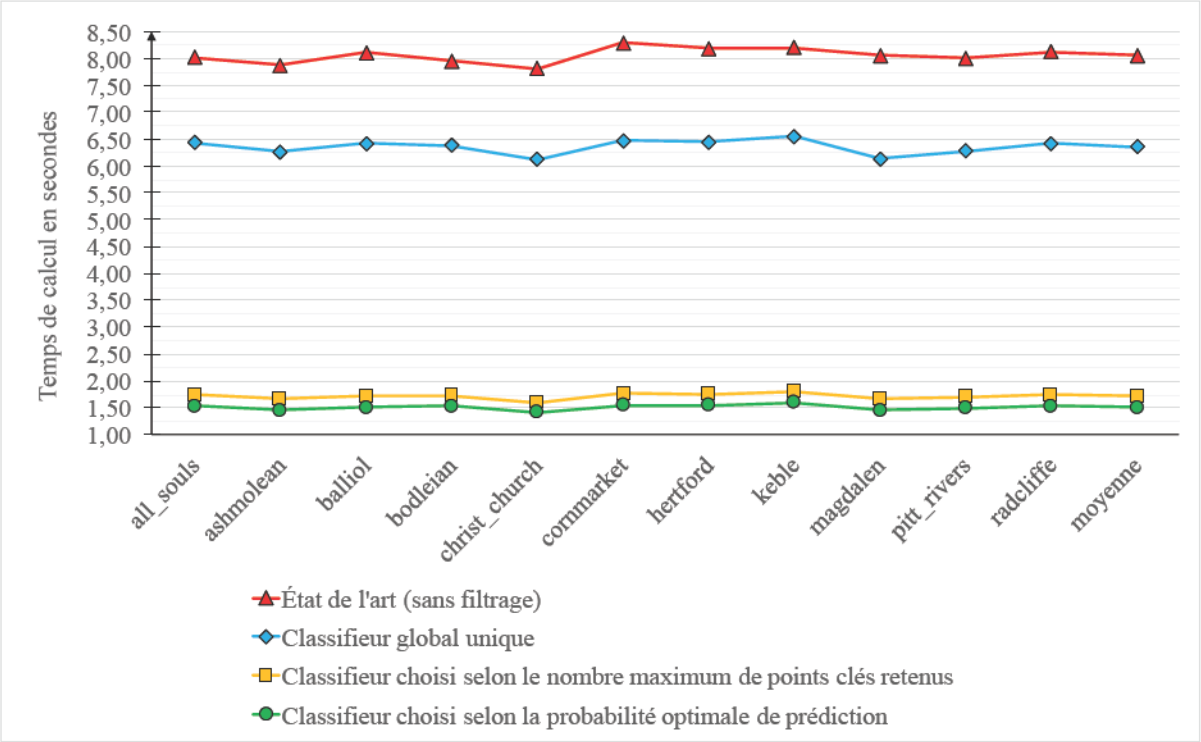


Figure 106. Temps de calcul moyen de recherches d'images par requête sur la base de données Oxford5k avec les différentes méthodes de filtrage des points d'intérêt.

Etudions maintenant si ces bons résultats sont également pertinents pour l'approche de représentation VLAD.

VII.5.2. Recherche d'images avec les descripteurs VLAD

Les résultats obtenus après filtrage multi-classifieur sont résumés dans le **Tableau 25**.

	Paris6k	**Oxford5k**
État de l'art (sans filtrage)	0,506	0,537
Classifieur choisi selon le nombre maximum de points clés retenus	0,527	0,549
Classifieur choisi selon la probabilité optimale de prédiction	0,540	N/A

Tableau 25. Résultats MAP de la recherche d'images dans la base de données Oxford5k avec la méthode VLAD après filtrage des points clés grâce à différents classifieurs SVM adaptés.

De même qu'avec la méthode BOW le filtrage obtenu avec différents classifieurs adaptés à chaque catégorie d'image et sélectionnés selon la probabilité de classification la plus élevée présente les meilleurs résultats *MAP*.

VII.6. Bilan

Les différentes expérimentations présentées dans ce chapitre confirment pleinement l'intérêt de filtrer les points d'intérêt pour n'en retenir qu'on sous-ensemble correspondant aux objets recherchés. Cette tendance est confirmée aussi bien pour les représentations par BOW que pour les descripteurs VLAD, pour les deux bases de données retenues Paris6k et Oxford5k. La classification des descripteurs suivant les classes *bâtiment* et *non-bâtiment* conduit à de meilleurs performances en terme de rappel et précision pour la recherche d'images en prenant en compte différents classifieurs SVM adaptés pour chaque catégorie de bâtiment de la base de données et en sélectionnant le plus approprié selon le critère de probabilité de classification optimale.

VIII. CONCLUSION ET PERSPECTIVES

Résumons à présent les principales contributions apportées et esquissons les verrous restant à débloquer dans la continuité de ces travaux.

VIII.1. Principales contributions

Dans cette thèse, nous avons présenté différents modèles permettant d'amélioration les méthodes de recherche d'images représentant de monuments/bâtiments dans des scènes urbaines. Le principe de l'approche proposée consiste à affiner/filtrer les points d'intérêt détectés sur une image de façon à ne conserver que les descripteurs représentant effectivement l'objet d'intérêt recherché et en éliminant le reste des points pouvant interférer et donc pénaliser le processus de recherche. De façon pragmatique dans notre contexte de recherche de bâtiment à partir de scènes publiques et urbaines, notre objectif est de conserver uniquement les descripteurs issus du bâtiment recherché et d'éliminer ceux provenant des bâtiments voisins, de la végétation, des véhicules, des piétons…

Cette sélection de points clés pertinents devant s'effectuer sur l'ensemble des images d'une base de données, le modèle défini se réalise de façon automatique après un entraînement supervisé de classifieurs binaires. Nous avons spécifié plusieurs modalités d'entraînement, en adoptant une technique de classification SVM.

La première méthode consiste à entraîner un unique classifieur indépendamment des différents bâtiments de façon à diviser les descripteurs en deux classes : ceux issus du bâtiment d'intérêt d'un côté (la classe *bâtiment*) et le reste de l'autre (la classe *non-bâtiment*). Une seconde méthode fait intervenir un ensemble de classifieurs SVM, avec un classifieur adapté pour chaque catégorie d'images définies dans la base de données.

Pour une image donnée, la classification des descripteurs suivant les classes *bâtiment* et *non-bâtiment* peut donc être prédite selon les différents classifieurs entraînés de façon adaptée pour chaque catégorie de la base de données. De façon à sélection le classifieur le plus approprié, nous avons donc défini deux mesures de pertinence d'un classifieur sur une image donnée. La première mesure est le nombre de descripteurs conservés après filtrage via ce classifieur, alors que la seconde concerne la probabilité de la prédiction moyenne faite par ce classifieur sur les différents descripteurs prédits comme appartenant à la classe *bâtiment*.

Grâce à ces deux mesures, nous avons par la suite pu définir un modèle de classification sémantique, c'est-à-dire d'associer une image inconnue à une catégorie d'images de la base de données de façon automatique.

D'autre part, dans le but de conserver une cohérence non seulement sémantique mais aussi visuelle, nous avons proposé une méthode de correction géométrique des prédictions de classification SVM en classes *bâtiment* et *non-bâtiment*. L'objectif est dans ce cas de corriger des points attribués à l'une ou l'autre des deux classes, isolés parmi des points attribués à l'autre classe. Pour ce faire, nous avons proposé une définition optimisée du voisinage d'un point par l'utilisation d'arbre de recherche quadtree. Pour chaque point clé l'influence de ses voisins permet donc de retrouver une cohérence géométrique et visuelle de la classification SVM des différents descripteurs.

VIII.2. Rappel des résultats

Dans nos différentes expérimentations, nous avons montré que l'utilisation de descripteurs adaptés conduit à de meilleures performances non seulement en termes de classification mais aussi pour la recherche d'image par similarité, selon les scores de rappel et de précision. Le critère le plus pertinent pour sélectionner un classifieur SVM approprié à une image donnée s'appuie sur la probabilité moyenne de la prédiction des différents attribués à la classe *bâtiment*. Cette méthode de filtrage a montré une amélioration significative des scores de rappel (de 13,39%) et de précision (de 12,25%) par rapport à la méthode initiale de l'état de l'art. Cela démontre de façon empirique la pertinence de la procédure de filtrage proposée.

D'autre part, nous avons pu remarquer que la méthode de correction géométrique par voisinage présente un plus faible taux de faux positifs dans notre classification SVM, c'est-à-dire qu'un moins grand nombre de points clés n'étant pas issus du bâtiment d'intérêt sont conservés par la suite.

VIII.3. Perspectives

Notre modèle proposé de correction géométrique de la classification SVM des descripteurs locaux ne permet pas de conduire à une amélioration des performances de recherche d'image et de classification. Cela s'explique par le fait que trop peu de points sont finalement conservés par la suite puisque souvent corrigés de façon à exclure les descripteurs non pertinents. Une approche retenant un plus grand nombre de points adéquats permettrait donc d'améliorer d'autant plus les mots clés pris en compte dans notre système. Cependant, il ne s'agit pas d'inclure des points clés de l'environnement autres que ceux issus du monument recherché, le but étant de retrouver un objet en particulier dans différentes images.

D'autre part, de façon à améliorer la précision de la classification sémantique, il serait nécessaire d'affiner les mesures de pertinence proposées d'un classifieur sur une image donnée. La seule information que nous ayons pu exploiter étant la distance signée de la prédiction par classification SVM, nous avons tout de même pu montrer que sélectionner les points clés pour ne tenir compte que des plus appropriés dans un système de recherche d'image permet d'en améliorer les performances de rappel et de précision.

BIBLIOGRAPHIE

[AlPP07] ALI, HAIDER ; PAAR, GERHARD ; PALETTA, LUCAS: Semantic Indexing for Visual Recognition of Buildings. In: *International Symposium on Mobile Mapping Technology*. Padua, Italy, 2007, S. 28–31

[Altm92] ALTMAN, N. S.: An introduction to kernel and nearest-neighbor nonparametric regression. In: *The American Statistician* Bd. 46 (1992), Nr. 3, S. 175–185 — ISBN 0003-1305

[AnTD05] ANDREASSON, HENRIK ; TREPTOW, ANDRE ; DUCKETT, TOM: Localization for Mobile Robots using Panoramic Vision, Local Features and Particle Filter. In: *International Conference on Robotics and Automation*. Barcelona, Spain : IEEE, 2005

[ArZi12] ARANDJELOVIC, RELJA ; ZISSERMAN, ANDREW: Three Things Everyone Should Know to Improve Object Retrieval. In: *Computer Vision and Pattern Recognition*. Washington, United States : IEEE Computer Society, 2012, S. 2911–2918

[ArZi13] ARANDJELOVIC, RELJA ; ZISSERMAN, ANDREW: All about VLAD. In: *Computer Vision and Pattern Recognition* : IEEE Computer Society, 2013, S. 1578–1585

[BAGN13] BELTRAN, ARTURO ; ABARGUES, CARLOS ; GRANELL, CARLOS ; NUNEZ, MANUELA ; DIAZ, LAURA ; HUERTA, JOAQUIN: A Virtual Globe Tool for Searching and Visualizing Geo-Referenced Media Resources in Social Networks. In: *Multimedia Tools and Applications*. Hingham, United States : Kluwer Academic Publishers, 2013, S. 171–195

[BaTG06] BAY, HERBERT ; TUYTELAARS, TINNE ; VAN GOOL, LUC: SURF: Speeded Up Robust Features. In: *European Conference on Computer Vision* Bd. 110 (2006), Nr. 3, S. 346–359

[Beza16] BEZAK, PAVOL: Building Recognition System Based on Deep Learning. In: *Artificial Intelligence and Pattern Recognition* : IEEE, 2016, S. 159–163

[Bied87] BIEDERMAN, IRVING: Recognition-by-Components: A Theory of Human Image Understanding. In: *Psychological Review* Bd. 94 (1987), S. 115–147

[BiMS09] BIORET, NICOLAS ; MOREAU, GUILLAUME ; SERVIERES, MYRIAM: Géolocalisation en milieu urbain par appariement entre une collection d'images et un SIG 2D. In: *Ingénierie des Systèmes d'Information* Bd. 14 (2009), Nr. 5, S. 107–131

[BoGV92] BOSER, BERNHARD ; GUYON, ISABELLE ; VAPNIK, VLADIMIR: A Training Algorithm for Optimal Margin Classifiers. In: *Computational Learning Theory*. Pittsburgh, United States : ACM, 1992, S. 144–152

[BoMG08] BOTTERILL, TOM ; MILLS, STEVEN ; GREEN, RICHARD: Speeded-Up Bag-of-Words Algorithm for Robot Localisation Through Scene Recognition. In: *International Conference Image and Vision Computing New Zealand*, 2008, S. 1–6

[BuZa13a] BURSUC, ANDREI ; ZAHARIA, TITUS: ARTEMIS @ MediaEval 2013: A Content-Based Image Clustering Method for Public Image Repositories. In: *ACM Multimedia*. Barcelona, Spain : CEUR Workshop Proceedings, 2013, S. 18–19

[BuZa13b] BURSUC, ANDREI ; ZAHARIA, TITUS: ARTEMIS at TRECVID 2013: Instance Search Task. In: *TRECVID Workshop*, 2013

[CAPZ15] CHATFIELD, KEN ; ARANDJELOVIC, RELJA ; PARKHI, OMKAR ; ZISSERMAN, ANDREW: On-the-Fly Learning for Visual Search of Large-Scale Image and Video Datasets. In: *International Journal of Multimedia Information Retrieval* : Springer, 2015, S. 75–93

[CDFW04] CSURKA, GABRIELA ; DANCE, CHRISTOPHER ; FAN, LIXIN ; WILLAMOWSKI, JUTTA ; BRAY, CEDRIC: Visual Categorization with Bags of Keypoints. In: *International Workshop on Statistical Learning in Computer Vision* : European Conference on Computer Vision, 2004, S. 1–22

[Chen13] CHEN, YANZHI: *Efficient and Robust Image Ranking for Object Retrieval*, University of Adelaide, 2013

[ChFr15] CHOI, JAEYOUNG ; FRIEDLAND, GERALD: *Multimodal Location Estimation of Videos and Images* : Springer Publishing Company, 2015

[ChHH09] CHUNG, YU-CHIA ; HAN, TONY ; HE, ZHIHAI: Building Recognition Using Sketch-Based Representations and Spectral Graph Matching. In: *International Conference on Computer Vision* : IEEE Computer Society, 2009, S. 2014–2020

[ChLi11] CHANG, CHIH-CHUNG ; LIN, CHIH-JEN: LIBSVM: A Library for Support Vector Machines. In: *ACM Transactions on Intelligent Systems and Technology* Bd. 2 (2011), Nr. 3, S. 27:1-27:27

[ChYZ14] CHEN, TAO ; YAP, KIM HUI ; ZHANG, DAJIANG: Discriminative Soft Bag-of-Visual Phrase for Mobile Landmark Recognition. In: *IEEE Transactions on Multimedia* Bd. 16 (2014), Nr. 3, S. 612–622

[CMPM11] CHUM, ONDREJ ; MIKULIK, ANDREJ ; PERDOCH, MICHAL ; MATAS, JIRI: Total recall II: Query Expansion Revisited. In: *Computer Vision and Pattern Recognition* : IEEE Computer Society, 2011, S. 889–896

[CORR16] CORDTS, MARIUS ; OMRAN, MOHAMED ; RAMOS, SEBASTIAN ; REHFELD, TIMO ; ENZWEILER, MARKUS ; BENENSON, RODRIGO ; FRANKE, UWE ; ROTH, STEFAN ; U. A.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: *Computer Vision and Pattern Recognition*, 2016, S. 3213–3223

[CORS15] CORDTS, MARIUS ; OMRAN, MOHAMED ; RAMOS, SEBASTIAN ; SCHARWACHTER, TIMO ; ENZWEILER, MARKUS ; BENENSON, RODRIGO ; FRANKE, UWE ; ROTH, STEFAN ; U. A.: The Cityscapes Dataset. In: *Computer Vision and Pattern Recognition* : IEEE, 2015, S. 3213–3223

[DGJP13] DELHUMEAU, JONATHAN ; GOSSELIN, PHILIPPE-HENRI ; JEGOU, HERVE ; PEREZ, PATRICK: Revisiting the VLAD Image Representation. In: *ACM Multimedia*. Barcelona, Spain : ACM, 2013, S. 653–656

[Diaz08] DIAZ, ALEJANDRO: Through the Google Goggles: Sociopolitical Bias in Search Engine Design. In: *Web Search : Multidisciplinary perspectives*. Bd. 14 : Springer Berlin Heidelberg, 2008 — ISBN 9788578110796, S. 11–34

[DSGS12] DOERSCH, CARL ; SINGH, SAURABH ; GUPTA, ABHINAV ; SIVIC, JOSEF ; EFROS, ALEXEI: What Makes Paris Look Like Paris? In: *ACM Transactions on Graphics* Bd. 31 (2012), Nr. 4, S. 101:1-101:9

[DuZZ15] DU, SHIHONG ; ZHANG, FANGLI ; ZHANG, XIUYUAN: Semantic Classification of Urban Buildings Combining VHR Image and GIS Data: An Improved Random Forest Approach. In: *ISPRS Journal of Photogrammetry and Remote Sensing* Bd. 105 (2015), S. 107–119

[EMCU02] EXTREMAL, MAXIMALLY STABLE ; MATAS, J ; CHUM, O ; URBAN, M ; PAJDLA, T: Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In: *British Machine Vision Conference* : BMVA Press, 2002 — ISBN 1-901725-19-7, S. 384–393

[FiBe74] FINKEL, RAPHAEL ; BENTLEY, JON: Quad Trees: A Data Structure for Retrieval on Composite Keys. In: *Acta Informatica* Bd. 4 (1974), Nr. 1, S. 1–9

[FiBo81] FISCHLER, MARTIN ; BOLLES, ROBERT: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. In: *Communications of the ACM* Bd. 24. New York, United States (1981), Nr. 6, S. 381–395

[FrSP06] FRITZ, GERALD ; SEIFERT, CHRISTIN ; PALETTA, LUCAS: A Mobile Vision System for Urban Detection with Informative Local Descriptors. In: *International Conference on Computer Vision Systems*. Los Alamitos, United States : IEEE Computer Society, 2006, S. 30

[GGHQ06] GROENEWEG, NIKOLAJ ; DE GROOT, BASTIAAN ; HALMA, ARVID ; QUIROGA, BERNARDO ; TROMP, MAARTEN ; GROEN, FRANCISCUS: A Fast Offline Building Recognition Application on a Mobile Telephone. In: *Advanced Concepts for Intelligent Vision Systems*. Antwerp, Belgium : Springer Berlin Heidelberg, 2006, S. 1122–1132

[GOSP13] GRONAT, PETR ; OBOZINSKI, GUILLAUME ; SIVIC, JOSEF ; PAJDLA, TOMAS: Learning and Calibrating Per-Location Classifiers for Visual Place Recognition. In: *International Journal of Computer Vision* Bd. 118 (2013), Nr. 3, S. 319–336

[GoTG04] GOEDEME, TOON ; TUYTELAARS, TINNE ; VAN GOOL, LUC: Fast Wide Baseline Matching for Visual Navigation. In: *Computer Vision and Pattern Recognition*. Los Alamitos, United States : IEEE Computer Society, 2004, S. 24–29

[HuMa05] HUTCHINGS, ROBIN ; MAYOL, WALTERIO: *Building Recognition for Mobile Devices: Incorporating Positional Information with Visual Features*. Bristol, UK, 2005

[IBPQ14] IONESCU, BOGDAN ; BENOIS-PINEAU, JENNY ; PIATRIK, TOMAS ; QUENOT, GEOREGES: Fusion in Computer Vision: Understanding Complex Visual Content. In: *Advances in Computer Vision and Pattern Recognition* : Springer Publishing Company, 2014, S. 29–52

[JaHa99] JAAKKOLA, TOMMI ; HAUSSLER, DAVID: Exploiting Generative Models in Discriminative Classifiers. In: *Advances in Neural Information Processing Systems*. Cambridge, United States : MIT Press, 1999, S. 487–493

[JaRa15] JAJAL, BRIJESH ; RAGHVENDRAN, VENKATESAN: Exploration of Real World Services by Google Goggles Freeware. In: *International Journal of Computer Science and Mobile Applications* Bd. 3 (2015), Nr. 2, S. 1–4

[JDSP10] JEGOU, HERVE ; DOUZE, MATTHIJS ; SCHMID, CORDELIA ; PEREZ, PATRICK: Aggregating Local Descriptors into a Compact Image Representation. In: *Computer Vision and Pattern Recognition*. San Francisco, United States : IEEE Computer Society, 2010, S. 3304–3311

[JeDS08] JEGOU, HERVE ; DOUZE, MATTHIJS ; SCHMID, CORDELIA: Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In: *European Conference on Computer Vision*. Berlin, Germany : Springer-Verlag, 2008, S. 304–317

[JGZY11] JI, RONGRONG ; GAO, YUE ; ZHONG, BINENG ; YAO, HONGXUN ; TIAN, QI: Mining Flickr Landmarks by Modeling Reconstruction Sparsity. In: *ACM Transactions on Multimedia Computing, Communications, and Applications* Bd. 7S (2011), Nr. 1, S. 31:1-31:22

[JLYL10] JOSHI, DHIRAJ ; LUO, JIEBO ; YU, JIE ; LEI, PHOURY ; GALLAGHER, ANDREW: Rich Location-Driven Tag Cloud Suggestions Based on Public, Community, and Personal Sources. In: *International Workshop on Connected Multimedia*. New York, United States : ACM, 2010, S. 21–26

[KaZB04] KADIR, TIMOR ; ZISSERMAN, ANDREW ; BRADY, MICHAEL: An Affine Invariant Salient Region Detector. In: *European Conference on Computer Vision*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2004, S. 228–241

[KhMW11] KHAN, NABEEL YOUNUS ; MCCANE, BRENDAN ; WYVILL, GEOFF: SIFT and SURF Performance Evaluation Against Various Image Deformations on Benchmark Dataset. In: *International Conference on Digital Image Computing: Techniques and Applications*. Washington, United States : IEEE Computer Society, 2011, S. 501–506

[KLCC06] KIM, YONGKWON ; LEE, KISUNG ; CHOI, KYUNGHO ; CHO, SEONG IK: Building Recognition for Augmented Reality Based Navigation System. In: *International Conference on Computer and Information Technology* : IEEE, 2006

[KnSP10] KNOPP, JAN ; SIVIC, JOSEF ; PAJDLA, TOMAS: Avoiding Confusing Features in Place Recognition. In: *European Conference on Computer Vision*. Heraklion, Greece : Springer, 2010, S. 748–761

[LBBH98] LECUN, YANN ; BOTTOU, LEON ; BENGIO, YOSHUA ; HAFFNER, PATRICK: Gradient-Based Learning Applied to Document Recognition. In: *Proceedings of the IEEE*, 1998, S. 2278–2324

[LHSA14] LI, JING ; HUANG, WEI ; SHAO, LING ; ALLINSON, NIGEL: Building Recognition in Urban Environments: A Survey of State-of-the-Art and Future Challenges. In: *Information Sciences* Bd. 277 (2014), S. 406–420

[LiAl09] LI, JING ; ALLINSON, NIGEL: Subspace Learning-Based Dimensionality Reduction in Building Recognition. In: *Neurocomputing* Bd. 73, Elsevier (2009), Nr. 1–3, S. 324–330

[LiAl13] LI, JING ; ALLINSON, NIGEL: Building Recognition Using Local Oriented Features. In: *IEEE Transactions on Industrial Informatics* Bd. 9 (2013), Nr. 3, S. 1697–1704

[LiCH09] LI, YUNPENG ; CRANDALL, DAVID ; HUTTENLOCHER, DANIEL: Landmark Classification in Large-Scale Image Collections. In: *International Conference on Computer Vision*. Kyoto, Japan : IEEE Computer Society, 2009, S. 1957–1964

[LiSh02] LI, YI ; SHAPIRO, LINDA: Consistent line clusters for building recognition in CBIR. In: *International Conference on Pattern Recognition* : IEEE Computer Society, 2002, S. 952–956

[Lowe04] LOWE, DAVID: Distinctive Image Features from Scale-Invariant Keypoints. In: *International Journal of Computer Vision* Bd. 60 (2004), Nr. 2, S. 91–110

[MiSc02] MIKOLAJCZYK, KRYSTIAN ; SCHMID, CORDELIA: An Affine Invariant Interest Point Detector. In: *European Conference on Computer Vision*. London, UK : Springer-Verlag, 2002, S. 128–142

[MiSc04] MIKOLAJCZYK, KRYSTIAN ; SCHMID, CORDELIA: Scale & Affine Invariant Interest Point Detectors. In: *International Journal of Computer Vision* Bd. 60 (2004), Nr. 1, S. 63–86

[MiSc05] MIKOLAJCZYK, KRYSTIAN ; SCHMID, CORDELIA: A Performance Evaluation of Local Descriptors. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* Bd. 27 (2005), Nr. 10, S. 1615–1630

[MPCM10] MIKULIK, ANDREJ ; PERDOCH, MICHAL ; CHUM, ONDREJ ; MATAS, JIRI: Learning a Fine Vocabulary. In: *European Conference on Computer Vision*. Heraklion, Crete, Greece : Springer Berlin Heidelberg, 2010, S. 1–14

[MTSZ05] MIKOLAJCZYK, KRYSTIAN ; TUYTELAARS, TINNE ; SCHMID, CORDELIA ; ZISSERMAN, ANDREW ; MATAS, JIRI ; SCHAFFALITZKY, FREDERIK ; KADIR, TIMOR ; VAN GOOL, LUC: A Comparison of Affine Region Detectors. In: *International Journal of Computer Vision* Bd. 65 (2005), Nr. 1–2, S. 43–72

[NiSt06] NISTER, DAVID ; STEWENIUS, HENRIK: Scalable Recognition with a Vocabulary Tree. In: *Computer Vision and Pattern Recognition*. Washington, United States : IEEE Computer Society, 2006, S. 2161–2168

[ObMa05] OBDRZALEK, STEPAN ; MATAS, JIRI: Sub-linear Indexing for Large Scale Object Recognition. In: *British Machine Vision Conferece* : BMVA Press, 2005

[OlTo01] OLIVA, AUDE ; TORRALBA, ANTONIO: Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope. In: *International Journal of Computer Vision* Bd. 42 (2001), Nr. 3, S. 145–175

[PaPS13] PANCHAL, PULKIT ; PANCHAL, SANDIP ; SHAH, SATISH: A Comparison of SIFT and SURF. In: *International Journal of Innovative Research in Computer and Communication Engineering* Bd. 1 (2013), Nr. 2, S. 323–327

[PCIS07] PHILBIN, JAMES ; CHUM, ONDREJ ; ISARD, MICHAEL ; SIVIC, JOSEF ; ZISSERMAN, ANDREW: Object Retrieval with Large Vocabularies and Fast Spatial Matching. In: *Computer Vision and Pattern Recognition*. Minneapolis, United States : IEEE Computer Society, 2007

[PCIS08] PHILBIN, JAMES ; CHUM, ONDREJ ; ISARD, MICHAEL ; SIVIC, JOSEF ; ZISSERMAN, ANDREW: Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In: *Computer Vision and Pattern Recognition* : IEEE, 2008

[PeDa07] PERRONNIN, FLORENT ; DANCE, CHRISTOPHER: Fisher Kernels on Visual Vocabularies for Image Categorization. In: *Computer Vision and Pattern Recognition* : IEEE Computer Society, 2007, S. 1–8

[PeSM10] PERRONNIN, FLORENT ; SANCHEZ, JORGE ; MENSINK, THOMAS: Improving the Fisher Kernel for Large-Scale Image Classification. In: *European Conference on Computer Vision*. Heraklion, Crete, Greece : Springer-Verlag, 2010, S. 143–156

[PhSZ11] PHILBIN, JAMES ; SIVIC, JOSEF ; ZISSERMAN, ANDREW: Geometric Latent Dirichlet Allocation on a Matching Graph for Large-Scale Image Datasets. In: *International Journal of Computer Vision* Bd. 95 (2011), Nr. 2, S. 138–153

[PISZ10] PHILBIN, JAMES ; ISARD, MICHAEL ; SIVIC, JOSEF ; ZISSERMAN, ANDREW: Descriptor Learning for Efficient Retrieval. In: *European Conference on Computer Vision*. Heraklion, Crete, Greece : Springer, 2010, S. 677–691

[Plat99] PLATT, JOHN: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: *Advances in Large Margin Classifiers* : MIT Press, 1999, S. 61–74

[PLSP10] PERRONNIN, FLORENT ; LIU, YAN ; SANCHEZ, JORGE ; POIRIER, HERVE: Large-Scale Image Retrieval with Compressed Fisher Vectors. In: *Computer Vision and Pattern Recognition* : IEEE Computer Society, 2010, S. 3384–3391

[Quin87] QUINLAN, J.R.: Simplifying decision trees. In: *International Journal of Man-Machine Studies* Bd. 27 (1987), Nr. 3 — ISBN 00207373 (ISSN)

[ScBS07] SCHINDLER, GRANT ; BROWN, MATTHEW ; SZELISKI, RICHARD: City-Scale Location Recognition. In: *Computer Vision and Pattern Recognition* : IEEE, 2007, S. 1–7

[Scha90] SCHAPIRE, ROBERT E.: The Strength of Weak Learnability. In: *Machine Learning* Bd. 5 (1990), Nr. 2, S. 197–227 — ISBN 0-8186-1982-1

[ShSG03] SHAO, HAO ; SVOBODA, TOMAS ; VAN GOOL, LUC: *ZuBuD - Zurich Buildings Database for Image Based Recognition*. Switzerland, 2003

[SiIt07] SIAGIAN, CHRISTIAN ; ITTI, LAURENT: Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* Bd. 29 (2007), Nr. 2, S. 300–312

[SiUn09] SIRMACEK, BERIL ; UNSALAN, CEM: Urban-Area and Building Detection Using SIFT Keypoints and Graph Theory. In: *IEEE Transactions on Geoscience and Remote Sensing* Bd. 47 (2009), Nr. 4, S. 1156–1167

[SiVZ13] SIMONYAN, KAREN ; VEDALDI, ANDREA ; ZISSERMAN, ANDREW: Deep Fisher Networks for Large-Scale Image Classification. In: *Advances in Neural Information Processing Systems*, 2013, S. 163–171

[SiZi03] SIVIC, JOSEF ; ZISSERMAN, ANDREW: Video Google: A Text Retrieval Approach to Object Matching in Videos. In: *International Conference on Computer Vision* : IEEE, 2003, S. 1470–1477

[SuFJ05] SULEIMAN, WASSIM ; FAVIER, ERIC ; JOLIVEAU, THIERRY: Buildings Recognition and Camera Localization Using Image Texture Description. In: *International Journal of Computer Vision* Bd. 61 (2005), Nr. 2, S. 159–184

[Suyk01] SUYKENS, J.A.K.: Nonlinear modelling and support vector machines. In: *IMTC 2001. Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Rediscovering Measurement in the Age of Informatics (Cat. No.01CH 37188)* Bd. 1 (2001), S. 287–294 — ISBN 0-7803-6646-8

[ToFW08] TORRALBA, ANTONIO ; FERGUS, ROB ; WEISS, YAIR: Small Codes and Large Image Databases for Recognition. In: *Computer Vision and Pattern Recognition*. Anchorage, United States : IEEE, 2008

[ToSP11] TORII, AKIHIKO ; SIVIC, JOSEF ; PAJDLA, TOMAS: Visual Localization by Linear Combination of Image Descriptors. In: *International Conference on Computer Vision* : IEEE, 2011, S. 102–109

[TrKJ08] TRINH, HOANG HON ; KIM, DAE NYEON ; JO, KANG HYUN: Facet-Based Multiple Building Analysis for Robot Intelligence. In: *Applied Mathematics and Computation* Bd. 205 (2008), Nr. 2, S. 537–549

[TuLo09] TURCOT, PANU ; LOWE, DAVID: Better matching with fewer features: The selection of useful features in large database recognition problems. In: *International Conference on Computer Vision Workshops* : IEEE, 2009

[Vapn98] VAPNIK, VLADIMIR N: Statistical Learning Theory. In: *Interpreting* Bd. 2 (1998) — ISBN 0471030031

[VeFe15] VEZHNEVETS, ALEXANDER ; FERRARI, VITTORIO: Looking Out of the Window: Object Localization by Joint Analysis of All Windows in the Image. In: *CoRR* (2015)

[ViJo01] VIOLA, PAUL ; JONES, MICHAEL: Robust Real-Time Object Detection. In: *International Journal of Computer Vision*, 2001

[VKMT13] VONDRICK, CARL ; KHOSLA, ADITYA ; MALISIEWICZ, TOMASZ ; TORRALBA, ANTONIO: HOGgles: Visualizing Object Detection Features. In: *International Conference on Computer Vision* : IEEE Computer Society, 2013, S. 1–8

[WeGo09] WEIZMAN, LIOR ; GOLDBERGER, JACOB: Urban-Area Segmentation Using Visual Words. In: *IEEE Geoscience and Remote Sensing Letters* Bd. 6 (2009), Nr. 8, S. 388–392

[WiBr07] WINDER, SIMON ; BROWN, MATTHEW: Learning Local Image Descriptors. In: *Computer Vision and Pattern Recognition*. Minneapolis, United States : IEEE Computer Society, 2007, S. 1–8

[WZLZ16] WANG, KEZE ; ZHANG, DONGYU ; LI, YA ; ZHANG, RUIMAO ; LIN, LIANG: Cost-Effective Active Learning for Deep Image Classification. In: *IEEE Transactions on Circuits and Systems for Video Technology* Bd. PP (2016), Nr. 99, S. 1–10

[YYGH09] YANG, JIANCHAO ; YU, KAI ; GONG, YIHONG ; HUANG, THOMAS: Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In: *Computer Vision and Pattern Recognition*. Miami, United States : IEEE Computer Society, 2009, S. 1794–1801

[ZaSh10] ZAMIR, AMIR ROSHAN ; SHAH, MUBARAK: Accurate Image Localization Based on Google Maps Street View. In: *European Conference on Computer Vision* : Springer Berlin Heidelberg, 2010, S. 255–268

[ZhKo07] ZHANG, WEI ; KOSECKA, JANA: Hierarchical Building Recognition. In: *Image and Vision Computing* Bd. 25 (2007), Nr. 5, S. 704–716

[ZhTM13] ZHANG, FAN ; TOURRE, VINCENT ; MOREAU, GUILLAUME: A General Strategy for Semantic Levels of Detail Visualization in Urban Environment. In: *Eurographics Workshop on Urban Data Modelling and Visualization*. Girona, Spain : The Eurographics Association, 2013, S. 33–36

[ZZSA09] ZHENG, YAN TAO ; ZHAO, MING ; SONG, YANG ; ADAM, HARTWIG ; BUDDEMEIER, ULRICH ; BISSACCO, ALESSANDRO ; BRUCHER, FERNANDO ; CHUA, TAT SENG ; U. A.: Tour the World: Building a Web-Scale Landmark Recognition Engine. In: *Computer Vision and Pattern Recognition*. Miami, United States : IEEE Computer Society, 2009

LISTE DES PUBLICATIONS

[PUBLIÉ] HASCOËT, NICOLAS ; ZAHARIA, TITUS: Building Recognition with Adaptive Interest Point Selection. In: *IEEE International Conference on Consumer Electronics*. Las Vegas, United States, 2017 — ISBN 9781509055449, p. 29-32

[ACCEPTÉ] HASCOËT, NICOLAS ; ZAHARIA, TITUS: Image Retrieval of Urban Landmarks with Filtered Local Information. In: *International Journal of Internet Technology and Secured Transactions*.

[ACCEPTÉ] HASCOËT, NICOLAS ; ZAHARIA, TITUS: Building Recognition with Adaptive Interest Point Selection. In: *International Symposium on Signals, Circuits and Systems*. Iasi, Romania, July 2017.