Discriminative image representations using spatial and color information for category-level classification

Rahat Khan

Ecole Doctorale ED488 “Sciences, Ingénierie, Santé”

Discriminative Image Representations Using Spatial and Color Information for Category-level Classification

Titre en Français:

Représentations discriminantes d'image intégrant information spatiale et couleur pour la classification d'image

Thèse avec label européen préparée par **Rahat Khan**
pour obtenir le grade de :

Docteur de l'Université Jean Monnet de Saint-Etienne
Domaine : **Informatique -Image- Vision**

Laboratoire Hubert Curien, UMR CNRS 5516
Faculté des Sciences et Techniques

Soutenance le 8 Octobre 2013 au Laboratoire Hubert Curien
devant le jury composé de:

Tinne Tuytelaars	Professeur, University of Leuven, Belgique.	Rapporteur
Theo Gevers	Professeur, University of Amsterdam, Pays-Bas.	Rapporteur
Gabriela Csurka	Chercheur, Xerox Research Centre Europe, Grenoble, France.	Examinateur
Joost Van de Weijer	Chercheur Senior, Computer Vision Center, Barcelone, Espagne.	Examinateur
Cécile Barat	Maître de Conférences, Université Jean Monnet, Saint Etienne, France.	Co-directeur de thèse
Damien Muselet	Maître de Conférences, Université Jean Monnet, Saint Etienne, France.	Co-directeur de thèse
Christophe Ducottet	Professeur, Université Jean Monnet, Saint Etienne, France.	Directeur de thèse

Acknowledgements

As much as I am loving to write acknowledgements as it comes with the sense of an accomplishment, I am afraid of being unjust to many people of being oblique about their support and help in the course of this thesis. Still, I am going to attempt the impossible to thank everyone who made this thesis possible.

At the very beginning, comes by advisers. Christophe Ducottet, an active adviser and a wonderful person, who have been very patient and supportive with me during my PhD work. The ever smiling Cécile Barat, who is equally nice and the one I depended on before every deadlines. Damien Muselet, my adviser from masters' thesis, who was the person to bug with the every little questions I had. And of course, Joost Van de Weijer, with whom I have passed some eureka moments. I am very grateful to him, for allowing me to work under him in CVC, Barcelona. Without these 4 persons, this thesis would have never been possible.

Next, I would like to thank all my colleagues in the Laboratory Hubert Curien who were there to discuss when I was stuck, lost or simply bored. A place which never felt like in a foreign land because of them. Also, my colleagues in CVC, Barcelona, specially, Fahad Khan who helped me in some of the experiments. I would also like to thank, everyone related to the administration of this PhD. The funding agency, the French embassy in Dhaka, the prefecture of Loire, the administration of the laboratory and the European Union who actually helped me to sow the seed of this PhD 5 years ago through the Erasmus Mundus scholarship.

My family, who are almost 10,000 kilometers away but no distance is far enough to keep them apart. Their advices, support and assurance have always kept me focused and relaxed. My friends, whom I miss dearly. Finally, I would like to thank my wife who stood by my side and kept an eye on all the other things so that I can concentrate on the PhD. She is the unmentioned co-author of every paper I have published during my PhD.

I did not mention many names in the fear of missing out some of the important ones. But if you know me and cared enough to read this, you should know that you are surely one of them.

Abstract

Image representation is in the heart of many computer vision algorithms. Different computer vision tasks (e.g. classification, detection) require discriminative image representations to recognize visual categories. In a nutshell, the bag-of-visual-words image representation is the most successful approach for object and scene recognition. In this thesis, we mainly revolve around this model and search for discriminative image representations.

In the first part, we present a novel approach to incorporate spatial information in the BoVW method. In this framework, we present a simple and efficient way to infuse spatial information by taking advantage of the orientation and length of the segments formed by pairs of similar descriptors. We introduce the notion of soft-similarity to compute intra and inter visual word spatial relationships. We show experimentally that, our method adds important discriminative information to the BoVW method and complementary to the state-of-the-art method.

Next, we focus on color description in general. Differing from traditional approaches of invariant description to account for photometric changes, we propose discriminative color descriptor. We demonstrate that such a color description automatically learns a certain degree of photometric invariance. Experiments show that the proposed descriptor outperforms existing photometric invariants. Furthermore, we show that combined with shape descriptor, the proposed color descriptor obtain excellent results on four challenging data sets.

Finally, we focus on the most accurate color representation i.e. multispectral reflectance which is an intrinsic property of a surface. Even with the modern era technological advancement, it is difficult to extract reflectance information without sophisticated instruments. To this end, we propose to use the display of the device as an illuminant while the camera captures images illuminated by the red, green and blue primaries of the display. Three illuminants and three response functions of the camera lead to nine response values which are used for reflectance estimation. Results show that the accuracy of the spectral reconstruction improves significantly over the spectral reconstruction based on a single illuminant. We conclude that, multispectral data acquisition is potentially possible with consumer hand-held devices such as tablets, mobiles, and laptops.

Résumé

La représentation d'image est au cœur de beaucoup d'algorithmes de vision par ordinateur. Elle intervient notamment dans des tâches de reconnaissance de catégories visuelles comme la classification ou la détection d'objets. Dans ce contexte, la représentation "sac de mot visuel" (Bag of Visual Words ou BoVW en anglais) est l'une des méthodes de référence. Dans cette thèse, nous nous appuyons sur ce modèle pour proposer des représentations d'image discriminantes.

Dans la première partie, nous présentons une nouvelle approche simple et efficace pour prendre en compte des informations spatiales dans le modèle BoVW. Son principe est de considérer l'orientation et la longueur de segments formés par des paires de descripteurs similaires. Une notion de "soft-similarité" est introduite pour définir ces relations intra et inter mots visuels. Nous montrons expérimentalement que notre méthode ajoute une information discriminante importante au modèle BoVW et que cette information est complémentaire aux méthodes de l'état de l'art.

Ensuite, nous nous focalisons sur la description de l'information couleur. Contrairement aux approches traditionnelles qui s'appuient sur des descriptions invariantes aux changements d'éclairage, nous proposons un descripteur basé sur le pouvoir discriminant. Nos expérimentations permettent de conclure que ce descripteur apprends automatiquement un certain degré d'invariance photométrique tout en surclassant les descripteurs basés sur cette invariance photométrique. De plus, combiné avec un descripteur de forme, le descripteur proposé donne des résultats excellents sur quatre jeux de données particulièrement difficiles.

Enfin, nous nous intéressons à la représentation de la couleur à partir de la réflectance multispectrale des surfaces observées, information difficile à extraire sans instruments sophistiqués. Ainsi, nous proposons d'utiliser l'écran et la caméra d'un appareil portable pour capturer des images éclairées par les couleurs primaires de l'écran. Trois éclairages et trois réponses de caméra produisent neuf valeurs pour estimer la réflectance. Les résultats montrent que la précision de la reconstruction spectrale est meilleure que celle estimée avec un seul éclairage. Nous concluons que ce type d'acquisition est possible avec des appareils grand public tels que les tablettes, téléphones ou ordinateurs portables.

Contents

Table des matières

List of Figures

List of Tables

Chapter 1

Introduction

The thesis, titled 'Machine Perception of Three Dimensional Solids' by Lawrence Gilman Roberts was published in 1961 from the Massachusetts Institute of Technology. It is regarded as the first attempt at the object recognition problem. More than half a century later, the problem still remains widely unsolved. In fact, the field of object recognition did not see much success until the rise of machine learning techniques.

Detecting and recognizing objects is thus one of the most important uses of vision systems in nature, and is consequently highly evolved. Indeed, humans can recognize more than 30,000 visual categories, and can detect objects in the span of a few hundred milliseconds. In recent past, machine based visual recognition has gained significant attention from the researchers. Automatic understanding of visual content has many applications, notably, surveillance, robotics, information retrieval, human computer interaction etc. The reason machine based object recognition is difficult is due to the large variation in the images. This variation could come from changes in viewpoint, illumination and scale. Moreover, deformation, occlusion, background clutter also contribute to the difficulty of the problem. Another major hurdle in object recognition is intra class variations inside a visual category. For examples, all instances of the visual category "chair" have four legs but their shape can vary a lot in function of their design (Figure 1.1). The human visual system has amazing capabilities of compensating for extreme changes in viewing conditions or intra class variations due to its ability to use complex prior knowledge acquired from long term learning. Although in case of computer vision, researchers are nowhere near human performance in this task, they have made considerable progress in the past few years.

The introduction of classification algorithms (e.g. support vector machine) has immensely contributed to the advancement of the category-level recognition. These algorithms take array of feature vectors characterizing images and their associated labels as input, to learn a set of classifiers. This

Figure 1.1: Some instances of the object category 'chair'.

is why, many computer vision methods focus on computing discriminative vector representation of images. In this context, this thesis proposes spatial and color information aware image representation. The main theme of the thesis is image representation, however, an important part is dedicated to image representation applied to category-level classification with bag-of-visual-words (BoVW) method. Specially, in the first part of the thesis, we work on the BoVW method to improve it using spatial information. This is why, we first explain briefly the BoVW method before presenting the background and motivation of this thesis.

1.1 The bag-of-visual-words method for object recognition

The idea of bag-of-visual-words (BoVW) is inspired by the bag-of-word method from the textual document processing domain, where documents are represented as histograms of textual words. In an analogy, distinct local image patches could be considered as visual words acting as building

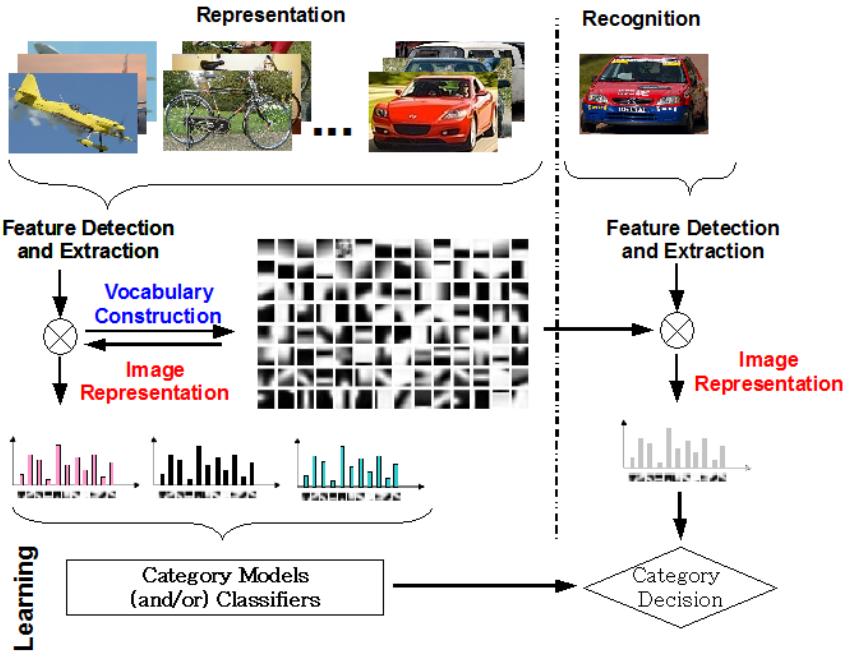

Figure 1.2: A pictorial depiction of the bag-of-visual-words framework.

blocks for natural images. The BoVW is one of the most successful object recognition method [50, 97]. It can be decomposed into four steps. The first one is keypoint detection. The goal of this step is to extract local regions from images. In the second step, features are computed from the local regions. During this step each local region is represented as a vector. Next, the obtained feature vectors are then quantized and termed as visual words. Finally, each image is represented as a histogram of occurrences of the visual words. This representation is known as the bag-of-visual-words. Figure 1.2 shows a pictorial depiction of the entire framework.

Although the BoVW method has been very successful in object/scene recognition, it has some shortcomings. For example, the BoVW representation is devoid of spatial information. Moreover, the feature quantization step results in the loss of discriminative power. Also, there are very few articles to integrate color information into image classification frameworks in general. Specially, there is a visible lacking of research to obtain better color description for discriminative tasks (e.g. classification) of the images. In this thesis, we target to look at discriminative image representation using spatial and color information.

1.2 Background and motivation

1.2.1 Spatial information for category-level image classification

As said before, one of the main drawbacks of the BoVW method is its inability to incorporate spatial information. Evidently, the histogram based representation does only provide frequency information where the spatial positions of the words are ignored. While frequency of visual features is important (for example, the visual category 'dog' is likely to have more 'textured' features than the category 'bottle'), for objects, spatial relationships among features can bring additional discriminative information. For example, spatial information can help to represent the global shape of an object which is not possible to obtain only using the BoVW representation. To this end, many methods have been proposed in the recent past. One of the methods, spatial pyramid representation (SPR) [50] has been very successful and shown to improve the classification accuracy significantly over the BoVW representation. SPR, proposed in the year 2006, has received extensive attention from the computer vision community. It has been cited for more than 2500 times. The idea of SPR is to divide the image into multiple sub-images using a simple grid. The division works in different levels where increase in level leads to finer grids (Figure 1.3). The BoVW representation is computed for each sub-image extracted from the grid. The final representation is obtained by a weighted concatenation of the BoVW representations of all the sub-images where the weights depend on the levels. Higher the level, finer the grid and higher the weights. Due to the simplicity of SPR and excellent performance on image classification tasks, it is used by default in many BoVW works. The success of SPR has led researchers to improve it in different ways [8, 30]. Nevertheless, these improvements often involve complex learning steps to learn different strategies to create sub-images or adapt weights on a validation set. Moreover, spatial pyramid only captures the global layout of the arrangements of the visual words in the image and which is by any means not the only spatial information. Having said that, it is not worthwhile to find a new spatial method to replace the SPR as it performs very well. Rather it is more advantageous to find spatial information complementary to the SPR and add it to the SPR. Recently, some authors have worked with this line of thought and obtained improved classification accuracy [99, 102]. It is worth mentioning that rather than using SPR, one can easily replace it with any other improved SPR techniques [8, 30] and then add the additional refined spatial information to further improve the accuracy.

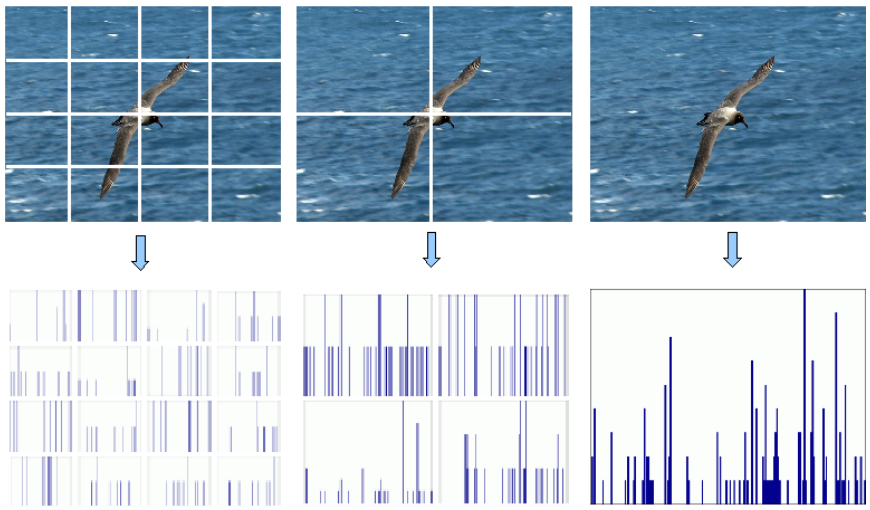

Figure 1.3: Spatial Pyramid Representation (SPR): the final histogram is obtained by concatenating all the histograms from individual regions.

1.2.2 Representation of color information

Color information plays an important role in recognition. The human visual system(HVS) is particularly good in using color to recognize objects. HVS is able to quickly segment an image using color information leading to superior understanding. Color also helps HVS to figure out the salient part of an image efficiently. Another amazing ability of HVS is color constancy. Color constancy ensures that the perceived color of objects remains relatively constant under varying illumination conditions. A green apple for instance, looks green to us at midday, when the main illumination is white daylight, and also at sunset, when the main illumination is reddish. HVS has certain degree of invariance to illumination changes while having excellent discriminative power. This balance in invariance and discrimination is desirable for many computer vision algorithms.
Although color is very important and one of the main cues used by HVS, efforts to exploit color information for category-level classification is largely ignored. Recently, Sande et al. [86] have shown that gradient based color features like SIFT performs well in this regard. On the other hand, Khan et al. [37, 38] have shown that pure color features used in conjunction with shape features can outperform gradient based color descriptors. There exist many different pure color descriptors [86, 87, 92]. Pure color descriptors directly deal with color components. They are often histograms of invariant color components (e.g. hue). The main objective of these descriptors remains to obtain invariance with respect to change in viewing conditions. To derive color invariant models a reflection model [49, 76] is almost always used. A reflectance model makes several assumptions and thus makes it dif-

ficult to generalize on real world settings. Recently, pure color descriptors have been used extensively for image classification [37, 38, 86]. However, as none of these descriptors are optimized for discriminative tasks, it is intuitive that they do not obtain optimal accuracy for discriminative tasks.
Formation of colors involves a material reflectance, an illumination and an observer. Reflectance is an intrinsic property of an object. It could be useful in material classification. Reflectance brings more information than trichromatic values and is invariant to viewing conditions. Many vision related applications could be highly benefited by reflectance based color representation. However, obtaining reflectance information of a material is technically difficult. It requires sophisticated instruments (e.g. spectro-photometers) which are very expensive. This is why multispectral representation of image for computer vision applications is not popular. There exist several works to facilitate multispectral imaging in low cost settings [70, 78]. However, many of them still require additional instruments(e.g. filters, light sources etc). The real challenge lies in finding solutions to multispectral reflectance acquisition problem with an affordable single device without additional equipments. In this thesis, we look at this aspect of color description as well.

1.3 Objectives and contributions

Above we discussed three aspects of spatial and color representation of images. This analysis has led us to the following three objectives of the thesis research.
To Enrich BoVW framework with spatial information: In the first part of the thesis, we focus on the infusion of spatial information into the BoVW framework. Spatial information is complex. It could be extracted from small regions(local) or the entire image(global). Additionally, it might be absolute or relative. We observe that, each spatial method usually deals with one particular type of spatial information. Thus, a single spatial method for BoVW is often not enough to take the maximum advantage of spatial arrangement of the visual words. However, as different spatial methods encode different information, they could very well be complementary to each other. Hence, we propose a new and simple spatial scheme for BoVW framework that performs well alone and which is complementary to the state-of-the-art spatial methods. To incorporate this information into the BoVW framework, we propose to compute spatial relationship among similar patches inside an image and encode this information into the BoVW representation. Our idea is motivated by that of [17, 77] where they show that self-similar patches provide discriminative information. This extension of the BoVW to integrate spatial information is presented in Chapter 3.
Discriminative Color Descriptors: Pure color descriptors are mostly

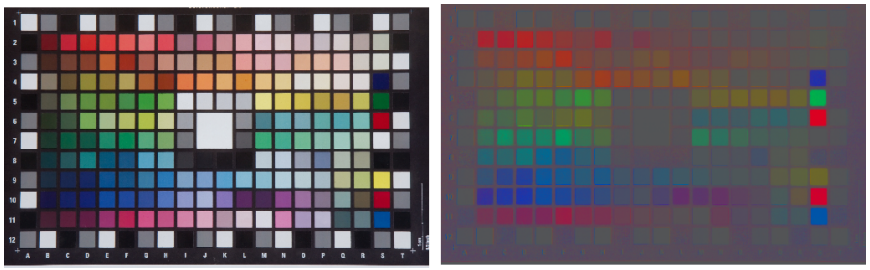

Figure 1.4: Loss of discriminative power due to invariance. On the left an original RGB image of a color checker under uniform illumination. On the right, invariance representation of the same image. Note the achromatic colors are not distinguishable anymore for the invariant image.

designed under the principles of color invariance. Invariance is always accompanied with the loss of discriminative power (Figure 1.4). Indeed, if multiple colors are mapped to a single color, which is what invariance does, those colors do not remain distinguishable. So, invariant color descriptors are not well suited for discriminative tasks. To this end, color descriptors optimized for higher discriminative power could be an interesting idea. In this direction, we present a method to learn a color descriptor given a classification problem. However, for generalization, a descriptor could be learned on a sufficiently large number of images. Although machine learning has been employed before to learn color invariance [4], to our knowledge, learning discriminative color descriptors is a new idea. Chapter 4 presents our proposal on discriminative color description with experimental results on multiple data sets.

Multi spectral data acquisition using handheld devices: Access to reflectance image of an object would be helpful for many computer vision applications. To this end, we propose to use handheld devices(e.g. laptops, tablets, smart phones) to obtain multispectral data of colored surfaces. Our method relies upon only the device itself and eliminates the need of any additional and expensive equipment. Accurate color communication is important for many industrial applications, to name a few, online shopping industry, make up industry and paint industry. Precise color communication with the clients are necessary for these industries. In our work, we enable the end clients with the power of multispectral reflectance acquisition which would make many applications plausible whereas they are currently not possible. We present our proposal regarding this topic in chapter 5.

In the next chapter, we discuss category-level visual recognition. Specifically, we present a comprehensive step-by-step overview of the BoVW method. We look at each step of BoVW method and discuss the state-of-the-art in the field.

Chapitre 1

Introduction

La thèse intitulée "Machine Perception of Three Dimensional Solids" de Lawrence Gilman Roberts a été publiée en 1961 au Massachsetts Institute of Technology. Elle est considérée comme la première contribution au problème de la reconnaissance d'objets. Plus d'un demi siècle plus tard, le problème reste largement ouvert. En fait, le domaine de la reconnaissance d'objet n'a pas connu de développements importants jusqu'à l'avènement des techniques d'apprentissage automatique.

Détecter et reconnaître des objets est l'une des plus importantes fonctions des systèmes de visions dans la nature et par conséquent elle est très évoluée. En effet, un homme est capable de reconnaître plus de 30000 catégories visuelles, et peut détecter des objets en quelques centaines de millisecondes. Récemment, la reconnaissance par vision artificielle a particulièrement retenu l'attention des chercheurs. L'analyse automatique du contenu visuel est utile dans de nombreuses applications notamment la surveillance, la robotique, la recherche d'information, l'interaction homme-machine, etc. La reconnaissance par vision artificielle est une tâche difficile à cause de l'importante variabilité des images. Cette variabilité peut venir d'un changement de point de vue, d'éclairement ou d'échelle. De plus, les déformations, occlusions, inhomogénéités d'arrière plan sont autant de difficultés supplémentaires. Un autre obstacle majeur en reconnaissance d'objet concerne les variations intra-classe au sein d'une catégorie visuelle. Par exemple, toutes les instances de la catégorie visuelle "chaise" possèdent quatre pieds, mais leur forme peut varier énormément en fonction de leur design (Figure 1.1)). Grâce à sa faculté à utiliser les connaissances a priori qu'il a acquises à partir d'un long apprentissage, le système visuel humain possède d'étonnantes capacités de compensation des ces changements importants de point de vue ou de ces variations intra-classe. Bien qu'en vision par ordinateur, les algorithmes soient encore très loin d'atteindre les performances humaines, la recherche dans ce domaine a fait des progrès considérables ces dernières années.

FIGURE 1.1 – Plusieurs instances de la catégorie "chaise".

La mise au point des algorithmes de classification (par exemple les machines a vecteurs de support) a contribué significativement aux progrès en reconnaissance de catégories d'images. A partir d'un tableau de vecteurs descripteurs caractérisant l'image et d'un jeu de labels associés, ces algorithmes apprennent un ensemble de classifieurs. L'une des difficultés en vision par ordinateur est d'être capable d'extraire des vecteurs descripteurs discriminants pour la représentation des images. Dans ce contexte, cette thèse se propose d'étudier des représentations d'image prenant en compte la couleur et les informations spatiales. Bien que le sujet principal soit la représentation d'image, une part importante de la thèse est consacrée aux représentations à partir de sacs de mots visuels (BoVW pour Bag of Visual Words) pour la classification. Notamment, dans la première partie de cette thèse, nous travaillons sur l'amélioration de la méthode des sacs de mots visuels par la prise en compte des informations spatiales. Ainsi, nous présentons d'abord brièvement cette méthode avant de présenter le contexte et les motivations de cette thèse.

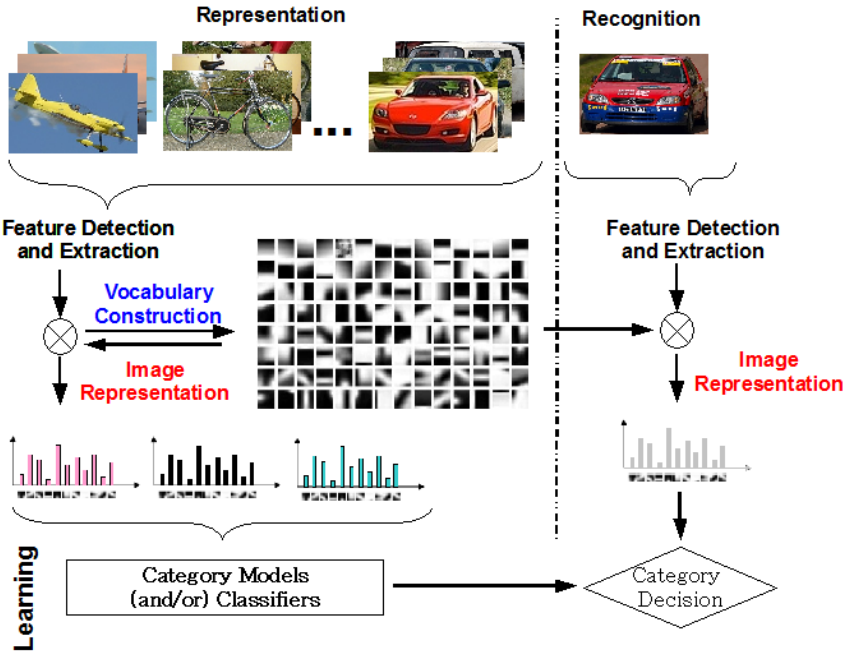

FIGURE 1.2 – Représentation schématique du modèle sac de mots visuels.

1.1 La méthode de sac de mots visuels pour la reconnaissance d'objets

L'idée des sacs de mots visuels (BoVW pour Bag of Visual Words) est inspirée des sacs de mots dans le domaine du traitement de documents textuels. Dans ce domaine, les documents sont représentés par des histogrammes doccurrence de mots textuels. Par analogie, de petits motifs locaux indépendants sur une image peuvent être considérés comme les mots visuels formant les constituants de base des images naturelles. Les sacs de mots visuels sont l'une des méthodes de reconnaissance les plus performantes [50, 97]. Elle peut être décomposée en quatre étapes. La première est la détection de points d'intérêts. Le but de cette étape est d'extraire des petites régions (ou motifs) de l'image. Lors de la deuxième étape, des caractéristiques (ou descripteurs) sont calculé à partir de ces petites régions, chaque région étant ainsi représentée par un vecteur. Ensuite, ces vecteurs descripteurs sont quantifiés pour définir les mots visuels. Finalement, chaque image est représentée comme un histogramme doccurrence des mots visuels. Cette représentation est connue sous le nom de "sac de mots visuels" ou bag of visual words en anglais. La figure 1.2 présente une illustration de cette représentation.

Bien que la méthode BoVW se soit montrée très performante en reconnaissance d'objet ou de scène, elle possède certain défauts. Par exemple

cette représentation ne prend pas en compte les informations spatiales. De plus, l'étape de quantification des descripteurs produit un perte de caractère discriminent. D'autre part, il y a très peu de travaux visant à intégrer l'information couleur dans le cadre de la classification d'image en général. Plus précisément, il y a un manque notable de recherches visant à obtenir de meilleurs descripteurs couleurs pour les tâches discriminantes telles que la classification d'images. Dans cette thèse, nous visons à proposer des représentations d'image discriminantes utilisant la couleur et les informations spatiales.

1.2 Contexte et motivations

1.2.1 Informations spatiales pour la classification d'image

Comme nous l'avons déjà dit, un des plus importants problèmes de la représentation BoVW est qu'elle nintègre pas d'informations spatiales. Bien évidemment, une représentation basée sur un histogramme ne contient qu'une information sur la fréquence sans considérer la position des mots visuels. Bien que cette information fréquentielle soit importante (par exemple la catégorie visuelle "chien" aura certainement plus de motifs texturés que la catégorie "bouteille"), pour les objets, les relations spatiales entre les descripteurs peuvent amener une information supplémentaire discriminante. Par exemple, l'information spatiale peut aider à représenter la forme globale d'un objet qui n'est pas possible à obtenir en utilisant seulement la représentation BoVW. Ainsi, beaucoup de méthodes ont été proposés récemment pour résoudre ce problème. L'une d'elle, la représentation en pyramide spatiale (SPR comme Spatial Pyramid Representation) [50] s'est révélée très performante en améliorant significativement le taux de classification comparativement à la représentation BoVW. La représentation SPR, proposée en 2006, a retenu particulièrement l'attention de la communauté scientifique en vision par ordinateur. Elle a été citée dans plus de 2500 publications. L'idée de la représentation SPR est de découper l'image en plusieurs sous-images en utilisant une grille simple. La division est effectuée à différents niveaux de telle manière qu'augmenter d'un niveau correspondent à une grille plus fine (Figure 1.3). La représentation BoVW est calculée pour chaque sous-image extraite de la grille. La représentation finale est obtenue par une concaténation pondérée des représentations BoVW issues de toutes les sous images. Le poids dépend du niveau de telle façon que les poids les plus forts sont attribués aux grilles les plus fines. Grâce à la simplicité du modèle SPR et de ses excellentes performances en classification d'image, il est utilisé par défaut dans beaucoup de travaux sur les sacs de mots. Ce succès a conduit les chercheurs à améliorer ce modèle dans différentes directions [8, 30]. Cependant, ces améliorations demandent souvent une étape d'apprentissage complexe pour apprendre différentes stratégies pour créer les

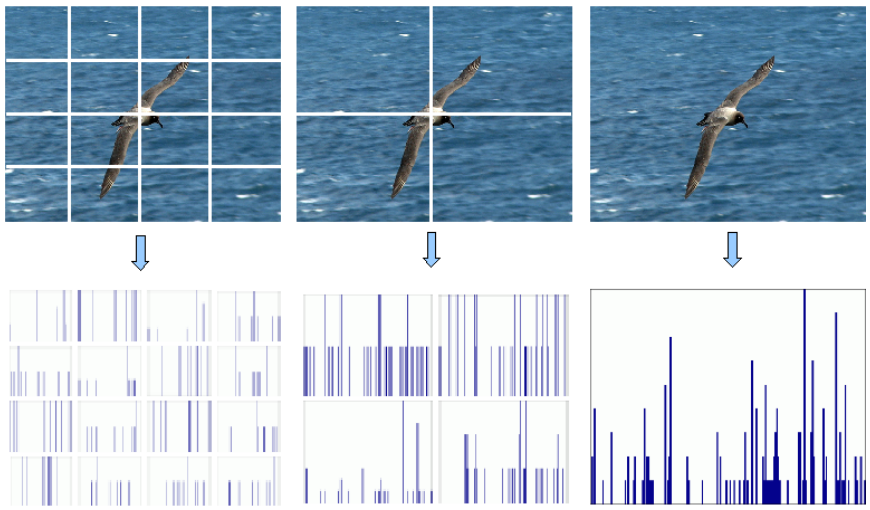

FIGURE 1.3 – Représentation en pyramides spatiales (Spatial Pyramid Representation, SPR) : l'histogramme final est obtenu en concaténant les histogrammes issus des différentes régions

sous-images ou pour adapter les poids sur un jeu de validation. De plus, la représentation SPR ne capture que la disposition globale des mots visuels dans l'image ce qui n'est pas la seule information spatiale intéressante. Néanmoins, ce n'est pas facile de trouver de nouvelles méthodes spatiales pour remplacer le modèle SPR car celui-ci est particulièrement performant. Il est certainement plus avantageux de chercher à extraire une information spatiale complémentaire et de l'intégrer au modèle SPR. Récemment, plusieurs auteurs ont travaillé dans cette direction et ont pu améliorer les taux de classification [99, 102]. Notons que, bien entendu, plutôt que d'utiliser le modèle SPR initial, on peut facilement le remplacer par n'importe quelle technique dérivée et ajouter ensuite une information spatiale élaborée pour améliorer encore les résultats.

1.2.2 Représentation de l'information couleur

La couleur joue un rôle important en reconnaissance. Le système visuel humain (SVH) sait particulièrement bien utiliser la couleur pour reconnaitre les objets. Il est capable de rapidement segmenter une image conduisant à une meilleure compréhension de celle-ci. La couleur aide aussi le système visuel humain pour détecter efficacement les parties saillantes d'une image. Une autre capacité étonnante du SVH est la constance des couleurs. Le terme constance des couleurs (en anglais color constancy) signifie que la couleur des objets est perçue de manière similaire indépendamment des conditions d'éclairement. Une pomme verte par exemple, apparaît verte à midi, alors qu'elle est éclairée en lumière blanche, et apparaît toujours verte au couché

de soleil, alors que la lumière est rougeâtre. Le SVH a un certain degré d'invariance aux changements d'éclairement tout en gardant un excellent pouvoir discriminant. Cet équilibre entre invariance et pouvoir de discrimination est souhaitable pour beaucoup d'algorithmes de vision par ordinateur.

Bien que la couleur soit importante et soit l'une des principales caractéristiques utilisées par le système visuel humain, peu de travaux ont été menés pour exploiter cette information pour la classification d'images. Récemment, Sande et al [86] ont montré que les descripteurs couleur basés sur le gradient tels que ceux dérivés du descripteur SIFT sont performants. D'un autre coté, Khan et al [37, 38] ont montré que les descripteurs couleur purs combinés avec des descripteurs de forme sont meilleurs que les descripteurs basés sur le gradient. Il existe beaucoup de descripteurs couleur purs [86, 87, 92]. Ils utilisent directement les composantes couleurs et sont souvent basés sur des histogrammes de ces composantes couleurs (par exemple la teinte). Leur principal objectif est d'être invariant par rapport aux changements de conditions d'observation. Pour obtenir l'invariance couleur, un modèle de réflectance [49, 76] est presque toujours proposé. Ce modèle est généralement basé sur des hypothèses difficilement respectées en conditions réelles. Bien que de nombreux travaux récents utilisent ces descripteurs couleur purs en classification d'images [37, 38, 86], aucun de ces descripteurs n'est optimisé vis à vis de son pouvoir discriminant. Il est donc plus que probable qu'il ne donne pas de résultats optimaux dans ce contexte de discrimination entre classes.

Le processus de formation des couleurs dépend de la réflectance du matériau, de l'éclairage et de l'observateur. La réflectance est une propriété intrinsèque de l'objet et peut être utilisée pour la classification des matériaux. Elle apporte plus d'information qu'un triplet de composantes couleurs et est invariante par rapport aux conditions d'observation. Beaucoup d'applications liées à la vision peuvent tirer parti d'une représentation couleur basée sur la réflectance. Cependant, accéder à cette information de réflectance d'un matériau est techniquement difficile. Cela demande des instruments sophistiqués (tels que les spectro-photomètres) qui sont très couteux. Ainsi, les représentations multispectrales pour les applications de vision ne sont pas très populaires. Il existe quelques travaux présentant des dispositifs d'imagerie multispectrale bon marchés [70, 78], mais beaucoup d'entre eux s'appuient sur des composants particuliers (comme des filtres, des sources de lumières,...). Le réel enjeu réside dans la recherche de solutions dacquisition multispectrale de réflectance à partir d'un seul système, à bas cot, ne nécessitant pas d'équipement additionnel. Dans cette thèse, nous nous intéressons également à cet aspect de la description de la couleur.

1.3 Objectifs et contributions

Nous avons discuté précédemment de trois aspects liés à la prise en compte de la couleur et des informations spatiales pour la représentation d'images. Cette analyse nous a conduits à définir trois objectifs pour notre travail de thèse.

Enrichir le modèle BoVW avec des informations spatiales : Dans la première partie de cette thèse, nous nous intéressons à l'ajout d'informations spatiales dans le modèle sac de mots visuels. Les informations spatiales sont de nature complexe. Elles peuvent être extraites à partir de petites régions (information locale) ou de l'image entière (information globale). Elles peuvent aussi être absolues ou relatives. Nous avons constaté que chaque méthode spatiale considérait généralement un type particulier d'information spatiale. Ainsi, une seule méthode spatiale pour le modèle BoVW n'est généralement pas suffisante pour prendre en compte la disposition spatiale des mots visuels. Par contre, comme les différents méthodes spatiales encodent une information différente, elles peuvent être très complémentaires les unes des autres. Ainsi, nous proposons une méthode simple et originale pour enrichir le modèle sac de mots par des informations spatiales. Cette méthode fonctionne bien seule et est aussi complémentaire aux autres méthodes de l'état de l'art. Son principe consiste à calculer les relations spatiales entre les motifs similaires d'une image et à encoder cette information dans la représentation BoVW. Notre idée est motivée par les travaux de [17, 77]qui montrent que les motifs auto-similaires apportent une information discriminante. Cette extension du modèle BoVW est présenté au chapitre 3.

Descripteurs couleur discriminants :

Les descripteurs couleur purs sont principalement conçus sur le principe de l'invariance couleur. L'invariance s'accompagne toujours d'une perte de pouvoir discriminant (Figure 1.4). En effet, si plusieurs couleurs sont fusionnées en une seule, ce qui est le principe de l'invariance, ces couleurs ne peuvent plus être distinguées. Ainsi, les descripteurs couleur invariants ne sont pas bien adaptés pour les tâches de discrimination. Il serait donc intéressant de disposer de descripteurs couleur optimisés par rapport à leur pouvoir de discrimination dans une tâche de classification. Nous proposons donc une méthode pour apprendre un descripteur couleur étant donné un problème de classification. Pour obtenir de bonnes performances en généralisation, ce descripteur devra être appris sur un nombre suffisant d'images. Bien que l'apprentissage automatique ait déjà été utilisé pour apprendre un invariant couleur [4], à notre connaissance, apprendre des descripteurs couleurs discriminants est idée originale. Le chapitre 4 présente notre proposition de description couleur discriminante avec des résultats

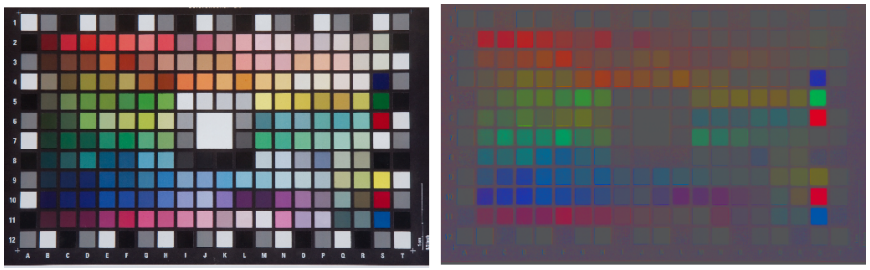

FIGURE 1.4 – Loss of discriminative power due to invariance. On the left an original RGB image of a color checker under uniform illumination. On the right, invariance representation of the same image. Note the achromatic colors are not distinguishable anymore for the invariant image.

expérimentaux sur plusieurs jeux de données.

Acquisition de données multispectrales à partir d'un appareil portable : L'accès à l'image de réflectance d'un objet est utile pour beaucoup d'applications de vision par ordinateur. Nous proposons d'utiliser un appareil portable (ordinateur portable, tablette, smart phone) pour extraire l'information multispectrale d'une surface colorée. Notre méthode s'appuie uniquement sur l'appareil portable sans nécessiter d'équipement additionnel onéreux. La transmission d'une information couleur précise est importante pour beaucoup d'applications industrielles comme par exemple le commerce en ligne ou l'industrie du maquillage et de la peinture. En permettant à un client d'acquérir et de transmettre simplement une information de réflectance multispectrale précise, notre méthode ouvre la porte à de nouvelles applications qui ne sont actuellement pas envisageables. Nous présentons cette méthode au chapitre 5.
Dans le chapitre suivant, nous nous intéressons à la reconnaissance visuelle de classes d'images. Plus précisément, nous détaillons toutes les étapes du modèle sac de mots visuels et nous présentons l'état de l'art actuel de ce domaine.

Chapter 2

Category-level Visual Recognition with the BoVW method

Résumé: Dans ce chapitre, nous donnons une description détaillée de la méthode de sac de mots. nous décrivons chaque étape de la méthode et de discuter de l'état de l'art pour chaque étape.

Abstract: In this chapter, we give a detail description of the BoVW method. We present a step by step review of the method. State-of-the-art practices for each step is also discussed.

2.1 Introduction

A category-level visual recognition system relates images in the real world to a visual category using models which are known a priori. Although the human visual system performs visual recognition effortlessly and instantaneously, for computers, this task is surprisingly challenging. This is due to the large variation in object poses and photometry which human visual system can cope with but algorithmic description of these variations on machines has been very difficult. Category-level image classification includes recognition of any visual concept including objects and scenes. The object recognition problem can be defined as a labeling problem based on models of known objects. Formally, given an image containing one or more objects of interest (and background) and a set of labels corresponding to a set of models known to the system, the system should assign correct labels to regions, or a set of regions, in the image.

There exist different approaches to handle the problem of category-level image classification. Nevertheless, bag-of-visual-words(BoVW) remains the

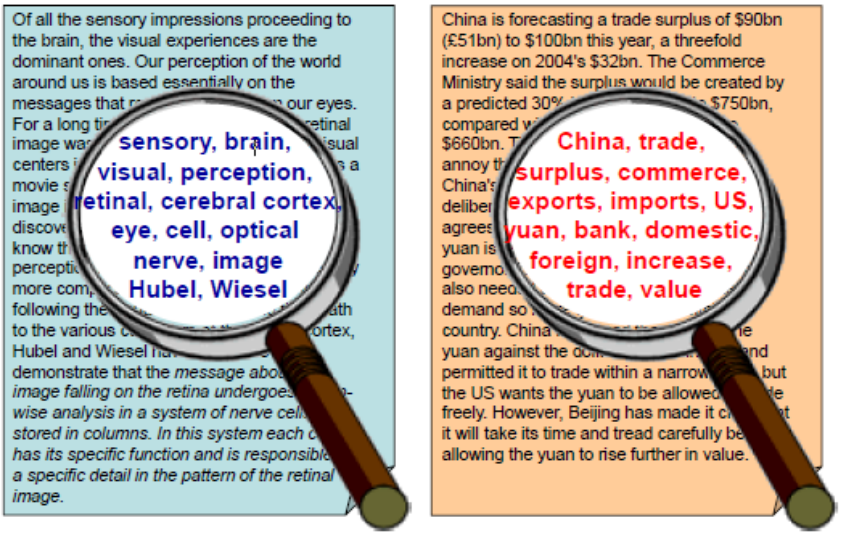

Figure 2.1: A intuitive depiction of the bag-of-words approach in text domain. The magnified words have the most frequent occurrence in the documents. Just by examining these words, these two documents could be classified in categories like 'neuroscience' or 'international trade'. Image courtesy of Silvio Savarese.

most successful method to accomplish this task. The idea of BoVW method originated from the language processing domain, where each document is represented as a histogram of words. Figure 2.1 shows how two different textual documents contains words from different semantic meaning and can help to recognize the category of the document itself.

Inspired by the success of BoVW in the language processing domain, the scientists in the computer vision domain extended this method to visual categories [15, 50].

In its simplest form, a bag of visual words is a histogram of quantized local features. The BoVW method comprises multiple steps. An image is duly represented as feature vectors and the final step usually employs a machine learning based classification algorithms (e.g. Random Forest, Support vector Machine). Each of these steps has been rigorously investigated by the researchers during the last decade. So, there exist a significant amount of literature related to each of them. In this chapter, we are going to provide a detailed overview of each stage of the BoVW pipeline.

2.2 Dissecting the bag-of-visual-words based classification pipeline

With very little exceptions, the BoVW method comprises of the following steps:

- Feature point detection
- Feature extraction
- Vocabulary construction
- Image representation

After the image representation step, different machine learning techniques are used to accomplish respective tasks(classification, detection etc.). Each of these steps has profound influence on the performance of the BoVW framework.

2.2.1 Feature point detection

The first stage within the BoVW approach involves detecting key points or regions in an image which are stable to affine changes. There exist multiple strategies for selecting regions in an image[63]. These strategies could be divided into multiple groups. The first group is known as interest point detectors. These types of detectors are more adapted to the textured scenes. They rely on finding salient points (such as corners, blobs etc.) in an image. Interest point detectors are often helpful for an object recognition task as they ignore the homogeneous areas and focus on the object and its surroundings in an image. Several interest point strategies have been proposed in the literature [63, 67]. Harris corner detector [31] is one of the very first detectors which is based on the idea of auto-correlation. An extension of the Harris corner detector, the Harris-Laplace point detector [63] focuses on locating corners that are scale invariant in an image. The Laplacian operator is used to find the scale of the corner. Another category of interest point detectors aim at detecting blob like structures in the image. Laplacian-of-Gaussian(LoG) [54] is a commonly used blob detector where an image is convolved using a gaussian kernel at certain scales to obtain a scale space representation. In other works, Lowe [57] uses difference-of-gaussian for scale invariant detection whereas Bay et al. [5] proposed to use hessian-laplacian to gain computational speed. Most of the existing interest point schemes make use of shape saliency as a selection criteria for detection. Another category focuses on regions rather than interest points. Maximally Stable Extremal regions (MSER)[61] falls into this category. MSER looks for connected components in a thresholded image (all pixels above/below a threshold) so that the region remains stable for a change of the threshold. MSER is more adapted to structured scenes. Finally, the dense sampling scheme is another type of interest region detector. Its region detection is not affected by the image content. Dense sampling is done by applying overlap-

Figure 2.2: Examples of some popular detectors applied on the same image. Top-left is the original image, top-right is the SIFT point detector, bottom-left is the laplacian of gaussian detector and the bottom-right is the dense detector.

ping grids on an image and considering every part of the image (very often in multiple scales). It is often advantageous for category-level classification where homogeneous regions can bring important information (e.g. scene classification). In figure 2.2, we present feature points detected by three popular feature point detectors.

2.2.2 Feature extraction

The next stage within the bag-of-words framework involves describing the extracted regions of an image. All the patches extracted in an image are normalized to standard size and descriptors are computed for all regions. Many features such as color, texture, shape have been used to describe visual information for object recognition. In the next paragraphs, we provide an overview of the two most commonly used visual cues namely, shape and color.

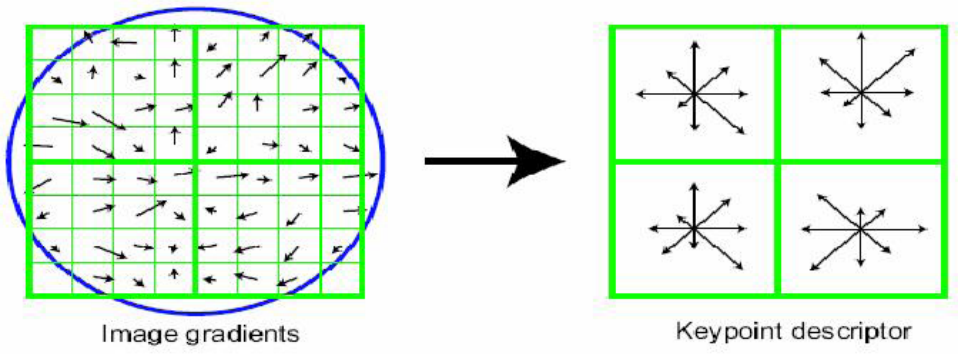

Figure 2.3: An example of SIFT computation. A region in an image is divided into four quadrants where each of the four quadrants contains 16 samples of the image gradient. The direction of the gradient together with magnitude samples are combined into a histogram of 8-bins gradient. Consequently, each of the four quadrants has its own histogram. The figure is taken from [56].

2.2.2.1 Shape feature extraction

The initial image descriptors were gaussian derivative-based and inspired by the human visual cortex [22, 45]. Soon after, gradient based image descriptors came to the scene and made a significant impact. At present, most of the image descriptors in use are gradient based [5, 57]. SIFT [57] is the most notable among the gradient based descriptors. In [62], the authors show that SIFT outperforms all the other descriptors in classification tasks. SIFT operates by computing gradients within a region of interest. The local appearance of the region is described by a gradient orientation histograms. The region of interest is first divided into a 4×4 grid of cells where each of the four quadrants has its own edge orientation histogram computed from the local gradient direction weighed by the magnitude of the gradient. The SIFT descriptor is highly invariant to changes in scale, illumination, and orientation. It is also partially invariant to 3D viewpoint change. Each SIFT key point has 132 dimensions where 128 are spatial orientation bins, plus the coordinates, rotation and the scale of the key point. Figure 2.3 shows computation of a SIFT descriptor in a region around a feature point detected by the interest point detectors.

Other than SIFT, there exist several other shape descriptors. Among them, SURF [5], PCA-SIFT [35], Daisy descriptors [83] are notable.

2.2.2.2 Color feature extraction

Color is a very important cue for category-level recognition, which can help increase the discriminative power of an object-recognition system and

also make it more robust to variations in the lighting and imaging conditions. However, color information comes with its own set of complexities. The RGB values of a digital color image do not only depend on the surface color but also on other factors like scene illumination, viewing geometry and surface gloss. To facilitate our discussion, we group the existing color descriptors in different categories described in the following paragraphs.

Gradient based color descriptors: In a recent study [86], Van de Sande et al. evaluated different color descriptors. They put particular emphasis on gradient based color descriptors. They employed SIFT on each channel of color images transformed into different color spaces and normalized images(e.,g. HSV-SIFT,HueSIFT, OpponnentSIFT, w-SIFT,rg-SIFT etc). They have shown that gradient based color descriptors perform very good on category-level recognition tasks. Particularly, opponent color space based SIFT descriptors outperform all the other color descriptors. Moreover, the authors have found that the recognition rates provided by the SIFT based color descriptors are much higher than those with the pure color descriptors(e.g. color statistics, color histograms). This indicates that, to obtain state-of-the-art accuracy, it is necessary to use shape information along with the color information to maximize discriminative power.

Invariant color descriptors: Invariant color descriptors are based on different assumptions on how light and matter interact. The Lambertian model [49], the dichromatic reflection model [76] and the Kubelka-Munk model [48] are the most popular surface reflection models. Invariant color descriptors are primarily inspired by the Human Visual System, which is color constant to some extent. There exist different invariant color descriptors. Each descriptor is hand designed to be invariant to one or few photometric changes. For example, color moments [64] is invariant to intensity shift, rg-histogram is invariant to light intensity change and hue-histogram [92] is invariant to intensity change and shift. There is always a trade-off between invariance and discriminative power. This is why, despite of their great theoretical background and motivation, invariant color descriptors often fail to obtain state-of-the-art classification accuracy.

Color names descriptor: Color names involve the assignment of linguistic color labels to image pixels. The 11 basic color terms of the English language are black, blue, brown, grey, green, orange, pink, purple, red, white and yellow [7]. Color names display a certain amount of photometric invariance because several shades of a color are mapped to the same color name. Van de Weijer et al. [87] learned color names from labeled images and used it as a color descriptor for image classification tasks. Along with a degree of photometric invariance, color names descriptor also allow to encode the achromatic colors such as black, grey and white, which are impossible to distinguish from the photometric invariance perspective. This leads the

color names descriptor to achieve higher discriminative power.

The limitations of color names descriptor are that, it is not optimized to be discriminative and it is not known how to extend beyond 11 color names as there is no known ordering after that.

Multispectral Color Representation: Reflectance is an intrinsic property of an object and it is the most accurate color description possible for a given object. Being able to capture multispectral reflectance of a scene would facilitate extremely precise color description. Human visual system has the capabilities to understand the intrinsic color of the object regardless of the viewing conditions. However, in case of computer vision systems, we are still far from this goal. Almost all the vision related systems are based on tri-chromatic color models. This is why multispectral color representation is not popular for category-level classification tasks.

2.2.3 Vocabulary construction

Feature extraction is followed by the visual vocabulary creation step. This step in BoVW methods is usually known as vector quantization. Generally, clustering algorithms are used to achieve this task in the descriptor space. Each cluster representative (typically the centroid) is considered as a visual word of the visual dictionary. Vocabulary construction could be unsupervised or supervised. The K-means clustering algorithm [86, 97] is the most common method to create such visual dictionaries even though other unsupervised methods such as K-median clustering [9], mean-shift clustering [34], hierarchical K-means [66], agglomerative clustering [51], radius based clustering [34], or regular lattice-based strategies [85] have also been used. One of the common features of these unsupervised methods is that they optimize an objective function fitting to the data but ignore the class information. The K-means algorithm minimizes the within-cluster sum of squares of distances. To create more discriminative visual words, one solution consists in using supervised approaches. In this context, some methods have been proposed to create class specific or concept specific multiple dictionaries [20, 72].
The quality of a visual dictionary depends on its size. Generally, improved results are obtained using larger visual vocabularies [97]. Also, with higher number of visual words in the vocabulary, the performance difference among different clustering algorithms slowly disappears. This is why, most of the recent work in this field employs the K-means algorithm.

2.2.4 Image representation

The next important step of BoVW framework is image representation. The baseline method is to compute a histogram of visual words (quantized

local features) and was introduced in [15]. Recent advances replace the hard quantization of features involved in this method with alternative encodings that retain more information about the original image features. This has been done in two ways: (1) by expressing features as combinations of visual words (e.g., soft quantization [25], local linear encoding [91]), and (2) by recording the difference between the features and the visual words (e.g., Fisher encoding [73], super-vector encoding [106]).

Combining Shape and Color Information: Generally, only shape/texture descriptors are used for local description for the BoVW framework. However, color can provide important information. Recently, [86] has shown that color information added to the shape descriptors can significantly improve classification accuracy. There are two main approaches to combine color and shape information, namely, early fusion and late fusion. In early fusion, the shape and color information are combined in the feature level and before the vocabulary construction step. In late fusion, the shape and color vocabularies are independently computed and the fusion is done by concatenating shape and color histograms computed independently. Often, shape and color information are independently learned from the training set. Recently, Khan et al. [38] introduced top down color attention based shape feature modulation to combine color and shape features to obtain state-of-the-art results.

2.2.5 Image classification

As explained in the previous section, the final image representation is a vector. Such vectors calculated from all the training images and their class labels are used as input to a machine learning algorithm for classification. A Support Vector Machine (SVM) is a learning algorithm typically used for classification problems . The goal of the SVM is to optimize "'generalization"', the ability to correctly classify unseen data. It addresses problems seen in other learning algorithms such as mistakes due to local minima, over fitting, and an inconveniently large number of tunable parameters. SVM maps training data in the "'input space"' into a high dimensional "'feature space"'. It determines a linear decision boundary in the feature space by constructing the "'optimal separating hyperplane"' distinguishing the classes with the help of the support vectors (Figure 2.4).

This allows the SVM to achieve a nonlinear (or linear depending on the type of mapping) boundary in the input space. The type of mapping is determined by the 'kernel function' in use. 'Kernel functions' also potentially help to avoid difficult computation in the feature space. The "'support vectors"' are those points in the input space which best define the boundary between the classes. The non-linear decision function in the original feature space is shown below:

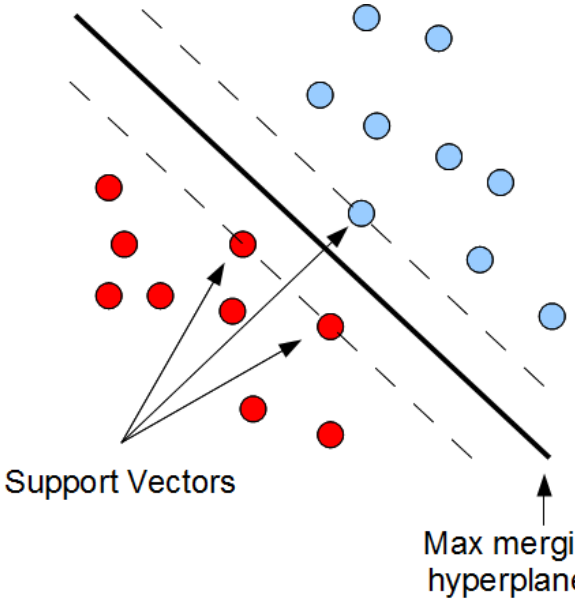

Figure 2.4: Classifier learnt using a linear support vector machine.

$$f(x_t) = \sum_i \alpha_i y_i K(\mathbf{x_i}, \mathbf{x_t}) + b \tag{2.1}$$

Here, x_i refers to the training sample, y_i is the corresponding training label, x_t is the test sample and K is the kernel function. Additionally, α_i and b are the parameters learned during training. SVM usually solves binary classification problems, where it finds a decision boundary to separate two classes from each other. However, it is possible to extend SVM to solve multiclass classification problem. It is generally done by dividing a multiclass problem into multiple binary classification problems. In particular, the most common technique in practice builds one-versus-rest classifiers (commonly referred to as "'one-versus-all"' classification) and chooses the class which classifies the test datum with greatest margin. Another strategy is to build a set of one-versus-one classifiers, and to choose the class that is selected by the most classifiers.
Over the last few years variety of different kernels have been proposed for SVM. For image classification, non-linear kernels like intersection or χ^2 have shown excellent performances. While χ^2 performs slightly better than intersection kernel, it requires tuning of an additional parameter and is slower than the intersection kernel [59]. In this thesis, we have used both kernels and a one-vs-all approach to multiclass SVM.

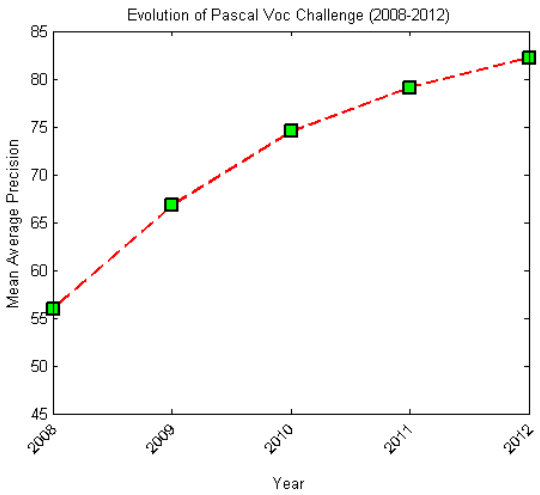

Figure 2.5: The evolution of the Pascal voc object recognition challenge. The classification accuracy is significantly increasing in a consistent matter each year.

2.3 Recent developments

In this section, we shortly discuss some of the recent advancement in category-level recognition. However, it is not remotely possible to do justice to all the works that has been published over the years attempting to solve the category-level recognition problem. Most of the recent developments in the field of category-level recognition is revolving around the BoVW method. Specially, there has been a lot of efforts to improve the final image representation from the local features [12, 73, 107]. The main idea is to incorporate higher order statistics than a simple visual word counting scheme. A significant amount of works has been focused on infusing the lost spatial information in the final BoVW representation [50, 55, 75]. Incorporating color information into the category-level recognition framework gained some interest [37, 38, 86]. A few works propose to unsupervised segmentation of image training sets into foreground and background in order to improve image classification performance [10, 11]. Also, method like Multiple kernel learning [65, 89] or SVM-KNN [103] has been successful by improving the learning step after image representation.
At last, we present the data obtained from Pascal VOC object recognition challenges to get an insight about the progress in this field. Pascal VOC object recognition challenge is a proper place to look for state-of-the-art advancement for object recognition as this challenge regularly took place for eight consecutive years and always involved the same categories of 20 objects collected from flickr photo streams with roughly the same difficulty level. To understand the evolution, we plot the mean of the maximum aver-

age precision obtained for each object for each challenge starting from 2008 until 2012. This plot gives a clear idea on the advancement made in the field of object recognition. The mean average precision rises impressively from 55.96% to 82.19% inside a time span of 5 years. Most of the teams in this contest employ the bag-of-visual-words method. To boost the performance the contestants usually apply multiple detectors and descriptors with advanced encoding techniques, higher order statistics, spatial pyramid and multiple kernel SVMs. The most recent winner of the challenge dealt with the problem of intra-class variation by inhomogeneous similarity aware subclass mining [13, 79].

2.4 Conclusion

In this chapter, we have reviewed the BoVW representation for image classification and object recognition. As mentioned, to improve classification accuracy, many methods have been proposed, using larger vocabularies, different encoding solutions or combining shape/color cues. However, none of these methods retain spatial relationship among the visual words in the image space. This is why, methods like spatial pyramid proposed by [50] is almost always used for image representation. To this end, this thesis proposes a way to add spatial information in the BoVW method. This method is presented in chapter 3. In chapter 4, we look at a more general problem of finding a color descriptor for discriminative task. We experimentally evaluated our descriptor using the BoVW method. The context of these problems and related works are discussed in the respective chapters.

Chapter 3

Spatial Information to Improve the BoVW Method

Résumé: Ce chapitre présente une nouvelle approche permettant d'ajouter de l'information spatiale dans la représentation sac-de-mots visuels pour améliorer la catégorisation ou la classification d'images. En effet, dans la représentation sac-de-mots traditionnelle, les vecteurs descripteur des images sont des histogrammes d'occurrences de mots visuels. Cette représentation est basée sur l'apparence et ne contient aucune information relative à la disposition des mots visuels dans le plan image 2D. Dans ce contexte, nous présentons une approche simple et efficace pour prendre en compte l'information spatiale. Notre approche vise une représentation explicite globale des relations spatiales entre mots visuels. A cette fin, nous exploitons l'orientation et la longueur des segments formés par des paires de motifs similaires. La similarité entre motifs est évaluée de manière souple par une pondération normale de la distance entre leurs descripteurs (soft similarity). Un histogramme normalisé d'angles-distances des paires est calculé, la contribution de chaque paire étant pondérée par la similarité. Un tel histogramme spatial est généré pour chaque type de mots visuels et tient compte de toutes les paires de motifs incluant ce mot visuel. Des expérimentations sur des bases de données standard connues ont prouvé que notre méthode est compétitive face aux autres techniques concurrentes. Aussi, il est montré que notre méthode apporte une information complémentaire à celle fournie par les pyramides et permet d'améliorer significativement les résultats de classification.

Abstract: This chapter presents a novel approach to incorporate spatial information in the bag-of-visual-words representation for category level and scene classification. In the traditional bag-of-visual-words method, feature vectors are histograms of visual words. This representation is appearance based and does not contain any information regarding the arrangement of

the visual words in the 2D image space. In this framework, we present a simple and efficient way to infuse spatial information. Particularly, we are interested in explicit global relationships among the spatial positions of visual words. Therefore, we take advantage of the orientation and length of the segments formed by pairs of similar patches. The similarity between patches are determined in a soft manner using a normal weighting depending on the distance among the patches in the descriptor space. An evenly distributed normalized histogram of angles and distances of the pairs is computed, where pairwise similarity weights are used. For each word type, we constitute one spatial histogram, which accounts for all pairs of patches involving that word type. Experiments on challenging data sets demonstrate that our method is competitive with the concurrent ones. We also show that, our method provides important complementary information to the spatial pyramid matching and can improve the overall performance. [1]

3.1 Introduction

In category level and scene classification, the BoVW method has shown excellent results in recent years [50, 53]. In this method, an image is represented as a histogram of quantized local features called visual words. However, being orderless, histogram representations do not preserve any spatial information. This is considered to be one of the major drawbacks of this very successful method.

Different methods have been proposed to incorporate spatial information into the BoVW representation [44, 50, 55, 75, 96]. Some of these approaches use spatial context during the vocabulary construction step to incorporate spatial information [74, 106]. Alternatively, the most popular approaches model the spatial arrangements of visual words on the 2D image space as an additional step [44, 46, 50, 82, 96, 100, 101, 102, 105]. These later methods are more popular as they obtain superior classification accuracies. It is due to the fact that they are able to capture both local and global relationships among the visual words.

In this context, the Spatial Pyramid Representation (SPR) [50] is probably the most notable work. Its principle relies in dividing the image into a sequence of increasingly coarser grids (eg. 4 by 4, 2 by 2 and 1 by 1) and computing a local BoVW histogram in each cell. Two images are then compared using an intersection kernel computed between the two corresponding sets of histograms. The success of SPR has drawn the attention of many researchers. As a result, several attempts to improve this method took place after it had been proposed. For example, Bosch et al. [8] have generalized the intersection kernel with other quasi-linear kernels like chi-

1. A part of this chapter appeared in the proceeding of the British Machine Vision Conference(BMVC), 2012 [40].

square and learned weights for each pyramid level rather than using fixed weights. Another method for weight learning of SPR was proposed in [30]. In another work, fisher vector based global spatial layout modeling [46] has shown excellent results as well.

Although SPR performs very well, it only captures the information about approximate global geometric correspondence of the visual words among images. That is why, many of the recent approaches propose to find features that are complementary to SPR [40, 99, 102]. In this context, some works consider relative spatial relationships between visual words as initially proposed by [75]. The principle is to build higher order statistics considering the occurrence of a given spatial configuration of visual words. Following the approach proposed by [55, 75], given a specific pair of visual words and a spatial neighborhood, the co-occurrence information can be encoded as a spatial histogram. Although this approach cannot perform as good as SPR when used alone, it was recently shown to provide significant improvement of the classification accuracy [40, 99, 102].

However, the abundance of visual words in an image makes it computationally expensive to explicitly model relative spatial relationship among visual words. Thus, methods like [55, 75] employ vocabulary compression or feature selection and model only local or semi-local spatial information to speed up the computation. Nevertheless, Elfiky et al. [19] have shown that vocabulary compression before spatial information extraction results into declined classification performance. In another work, Parikh [69] examines the human vs machine performance on jumbled images and concludes that existing machine vision techniques are already effective in modeling local information from images, thus future research efforts should be focused on more advanced modeling of global information.

Based on these observations, in this work, we propose a way to model the global and local relative spatial distributions of visual words. We compute pairwise spatial histograms to capture the global distribution of similar patches. However, to identify similar patches we do not impose any hard threshold, rather we introduce the notion of soft-similarity between two image patches. Each pair of patches gets a similarity score according to their distance in the descriptor space. To compute the pairwise spatial histograms, for a given visual word and all the pairs of patches, we consider those where at least one of the members belongs to that visual word. Then, we consider the orientation and length of the segment formed by each pair in the image space and calculate a normalized spatial angle-distance histogram. The soft-similarity score is used to weight the contribution of each pair to the spatial histograms.

Note that our method eliminates a number of drawbacks from the previous approaches by i) adopting a simpler word selection technique that supports fast exhaustive spatial information extraction ii) enabling infusion of both local and global spatial information iii) being robust to translation

that often occurs in object classification context.

The rest of the chapter is organized in the following way: the next section describes a review of the related works. Section 3.3 presents our approach to incorporate spatial information into the BoVW representation. Section 3.4 describes the implementation details and section 3.5 presents the results on different benchmarks and comparisons with several other methods. Section 3.6 concludes the article pointing towards our future works.

3.2 Related works

In the previous section, we have briefly mentioned some works that successfully attempted to incorporate spatial information into the BoVW representation. In this section, we are going to detail some of those methods. We are specially going to focus on the methods that model relative layout of the visual words [55, 75] as our method also models similar information. Along with them, we discuss some other global methods like SPR [50], fischer vectors [46] and methods that improve SPR infusing additional spatial information [99, 102]. All the methods described in this section will be compared in the experimental section.

Correlograms [33] have been widely used in texture classification and image indexing. Recently, they have been successfully utilized to capture information on spatial interaction among visual words [75, 105]. Savarese et al. [75] have employed correlograms to model relationships among the visual words and achieved improved classification accuracy. This approach captures the relationship between visual words as a function of distance in the image and constructs a correlogram matrix. The authors call each element of this matrix a correlaton and compute histograms of correlatons as the final representation.

An improvement in the results is shown by an augmentation of the correlaton representation with the regular bag-of-word representation. As correlogram is a function of distances, choices of those distances influence the spatial information captured and also the classification accuracy. The authors have argued that small and large distances should be considered to cover the whole image thus take into account both the global and local interaction of the visual words. In this case, efficient algorithms must be used to cut down computational time. In another work, Liu et al. [55] introduce an integrated feature selection and spatial information extraction technique to reduce the number of features and also to speed up the process. The process of feature selection and second order feature(e.g. spatial information) extraction are run alternatively at each iteration of the algorithm. At each round, feature selection selects one feature, and feature extraction pairs this feature with each of the previously selected features. As in [75], the final feature is constructed by the concatenation of the first and second order

features. Note that all of the previous methods under this category only deal with local and semi-local information, although global spatial methods described in the next paragraph very often outperform the local ones.

In the case of global spatial information, SPR [50] has shown very good performances on many challenging image datasets. SPR combines aggregated statistics of fixed subregions from different levels of a pyramid. Interestingly, this method is not invariant to global geometric transformations. Moreover, SPR lacks information about relative positions of visual words and local spatial patterns. Zhang et al. [102] use different heuristic approaches with success to infuse additional spatial information into the SPR and Yang et al. [99] use co-occurrence information to improve it. We conclude the related works section citing one of our previous work on this very problem. In [40], we have shown that global spatial orientations of intra-type visual words are discriminative and significantly improve the classification accuracy. Our current proposal could be considered as a step forward of that work. In this work, we extend the spatial encoding of BoVW to intra and inter type visual word using the notion of soft-similarity and include spatial distance with orientation. In the next section, we are going to introduce our notations with the brief explanation of the BoVW method.

3.3 Encoding distance-orientations information of similar patches

The principle of our method is to use pairwise spatial histograms to encode spatial co-occurence of similar word pairs. In this section we first present original pairwise spatial histograms introduced by Liu et al [55], then we introduce pairwise spatial histograms of similar patches and finally our image representation.

3.3.1 Pairwise spatial histograms

In the BoVW method, each image I is represented in terms of image descriptors [57, 95]:

$$I = \{d_1, d_2, d_3, d_4, \dots d_N\} \tag{3.1}$$

where d_i is a vector which is the description of the image patch i and N is the total number of patches in the image. Typically, K-means unsupervised clustering is applied on a large set of descriptors obtained from the training images to compute cluster centers $W = \{w_1, w_2, w_3, w_4 \dots w_K\}$ called visual vocabulary, where K is the predefined number of clusters. Each patch i of the image is then assigned to a visual word $w(d_i)$ which corresponds to the nearest center in the descriptor space. If a soft assignment model or a sparse coding is used, the corresponding visual word is the one having the highest weight.

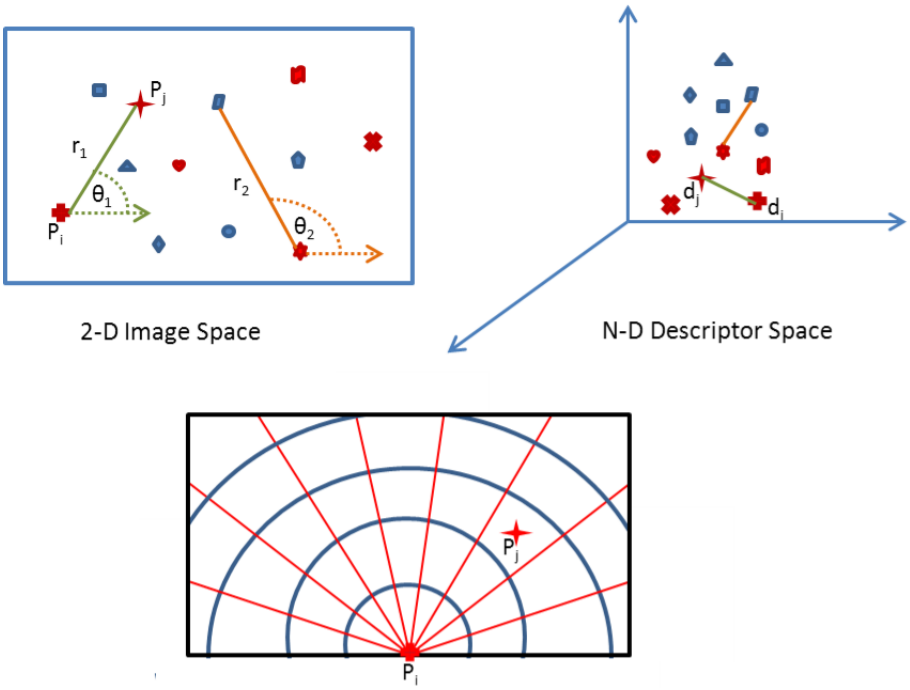

Figure 3.1: In this figure, each shape represents a different descriptor and all the descriptors with the same color belongs to one particular visual word. To encode spatial information, we use the distance and orientation information between pairs of patches in the image space (top-left) as well as their distance in the descriptor space(top-right). We consider inter and intra type word based on their proximity in the descriptor space. At the bottom, discretization of the image space used to define spatial histograms. Translating reference patch P_i (resp. P_j) at the center, the position of patch P_j (resp P_i) gives the bin number.

In Liu et al [55], a pairwise spatial histogram is defined according to a discretization of the spatial neighborhood into several bins encoding the relative spatial position (distance and angle) of two visual words (Figure 3.1). Given a specific pair of patches (P_i, P_j), it is defined as the count of all occurrences of P_j falling into a specific spatial bin relatively to P_i, the count being averaged for all instances of P_i.

3.3.2 Motivation of considering similar cues

The number of possible pairs of visual words is potentially very large and thus, Liu et al proposed to select only discriminative pairs. In [40], we have proposed another alternative which is to consider only pairs of identical visual words. The motivation came from the previous works [17, 77]

Figure 3.2: Discriminative power of spatial distribution of intra type visual words. Four images from Caltech101 dataset are shown. The black squares refer to identical visual words across all the images. For the two motorbikes in the left, the global distribution of the identical visual words is more similar than the ones in Helicopter or Bugle image.

where the authors have argued that modeling the distribution of similar cues across an image can give discriminative information about the content of that image. Figure 3.2 shows an example which gives an intuition to better understand the idea. In this illustration, we consider patches associated with the same visual word as similar cues.

3.3.3 Pairwise spatial histograms of similar patches

The notions of similar cues and similar words are not equivalent. If we consider clusters delimiting visual words in the descriptor space, two cues at the cluster borders could be very similar being in different clusters. Similar cues in this context are more related to a small inter-patch distances in the descriptor space.

Hence, we propose to analyze the spatial positions of the patches which are situated in proximity in the descriptor space. To avoid the use of a threshold to identify similar patches (hard similarity), we consider all the pairs of patches and we weight the contribution of each pair as a decreasing function $g(x)$ of their distance x in the descriptor space (soft similarity). We propose to use a gaussian function of standard deviation σ defined by:

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \tag{3.2}$$

This parameter gives us the control to highly weight patches that are in close proximity in the descriptor space and to ignore the ones which are far. More information about the choice of this parameter can be found in section 3.4.2. More formally, we consider the set S_k of all the pairs of patches where at least one patch in the pair belongs to the visual word w_k. A given pair $(P_i, P_j) \in S_k$ is characterized both by a pair of descriptors (d_i, d_j) and a pair of positions in the image space denoted (p_i, p_j) (Figure 3.1). A pairwise spatial histogram of similar patches is then defined considering a discretization of the image space into M bins with an angle $\theta \in [0, \pi[$ split into M_θ equal angle bins and the radius $r \in [0, R]$ split into M_r radial bins so that $M = M_\theta M_r$. For example, in the illustration at the bottom

of figure 3.1, the total number of bins M=45 (M_θ=9 and M_θ=5). The values of M_θ and M_r will be determined in the experimental section and the maximum radius R is chosen to be the diagonal of the image, in order to reduce scale sensitivity.

The bin count $H(m)$ of the spatial histogram of similar pairs H_k corresponding to the visual word w_k is then given by:

$$H_k(m) = \sum_{(P_i,P_j)\in S_k} g\left(|d_i - d_j|_2\right) \mathbb{1}_{\text{bin}(m)}(p_j - p_i) \tag{3.3}$$

where $|d_i - d_j|_2$ is the ℓ^2 distance in the descriptor space and $\mathbb{1}_{\text{bin}(m)}$ is the indicator function of bin m such that:

$$\begin{cases} \mathbb{1}_{\text{bin}(m)}(v) = 1 & \text{if } v \in bin(m) \\ \mathbb{1}_{\text{bin}(m)}(v) = 0 & \text{otherwise} \end{cases} \tag{3.4}$$

Note that, due to symmetric considerations, angle bins are discretized in the $[0, \pi[$ interval. Thus, given a pair of positions (p_i, p_j) the corresponding histogram bin is determined taking either p_i or p_j as reference point which is equivalent to consider either $(p_j - p_i)$ or $(p_i - p_j)$ vectors (figure 3.1). To evaluate the benefit of using inter-patch distance in the descriptor space, we will also consider a pairwise spatial histogram defined with identical visual words. Formally, if S_k^* denotes the set of all the pairs of patches for which both patches belong to visual word w_k, the hard pairwise histogram H_k^* is defined as:

$$H_k^*(m) = \sum_{(p_i,p_j)\in S_k^*} \mathbb{1}_{\text{bin}(m)}(p_j - p_i) \tag{3.5}$$

For a visual word w_k, equation 3.3 can be used to compute the entire histogram H_k. We propose the way to combine all H_k resulting from the different words in section 3.3.4.

3.3.4 Image representation

As explained in the previous section, for each visual word w_k, we obtain one spatial histogram. This histogram encodes spatial information (distance and orientation) of pairwise similar patches (intra and inter type visual words), where at least one of the patches belongs to w_k. This modularity facilitates simple way to assemble the spatial histograms and to obtain the final representation. We define three different representations: the soft pairwise similarity angle-distance histogram SPS_{ad} derived from the classical BoVW histogram, $SPS_{ad}+$ its combination whith SPR, and $SPS_{ad}^{1800}+$ a more compact version.

3.3.4.1 Soft Pairwise Similarity angle distance histogram SPS_{ad} representation

To obtain SPS_{ad} representation from the classical BoVW histogram, we use a 'bin replacement' technique. Bin replacement literally means to replace each bin of the BoVW frequency histogram with the spatial histogram H_k associated to w_k. The sum of all the bins of the spatial histogram obtained from one visual word w_k is normalized to the number of occurrences of this word in the whole image. By this way, we keep the frequency information intact and add the spatial information. The dimensionality of our representation $S = N \times M$ depends on the vocabulary size(N) and the number of angle-distance bins of the spatial histogram (M).
On the other hand, if we only consider hard pairwise spatial histograms H_k^* instead of soft pairwise spatial histograms H_k, we obtain a hard pairwise similarity histogram, denoted as HPS_{ad}.

3.3.4.2 Combination of SPS_{ad} with SPR

The SPS_{ad} representation is complementary to local first order BoVW representations as SPR, we propose to combine SPS_{ad} with SPR. We take the finest level of a 2-level SPR representation and concatenate it with SPS_{ad} without any weights, we denote this representation as $SPS_{ad}+$. Indeed, the dimensionality of the $SPS_{ad}+$ representation is the sum of that of the SPR and the SPS_{ad} representation. For a vocabulary size of N and a 2-level SPR, this dimensionality is $N \times (16 + M)$, 16 being the total number of local histograms in a 2-level SPR.

3.3.4.3 Dimensionality reduction

One of the drawbacks of the $SPS_{ad}+$ compared to SPR is the high dimensionality of the feature vectors. As compact representation is desirable, we use feature clustering to obtain a compact representation of the $SPS_{ad}+$ descriptor as shown in Elfiky et al. [19]. In their work, Elfiky et al. [19] employ divisive information theoretic clustering(DITC) proposed by Dhillon et al. [18] to compress the SPR representation without significant loss of discriminative information. The DITC algorithm minimizes the within-cluster Jensen-Shannon divergence while simultaneously maximizing the between-cluster Jensen-Shannon divergence. We compress our feature vectors down to 1800 dimensions. We denote this representation as $SPS_{ad}^{1800}+$.

3.4 Experimental protocol

In this section, we present the data sets used and the implementational details. We will evaluate different aspects of the SPS_{ad} representation for image classification.

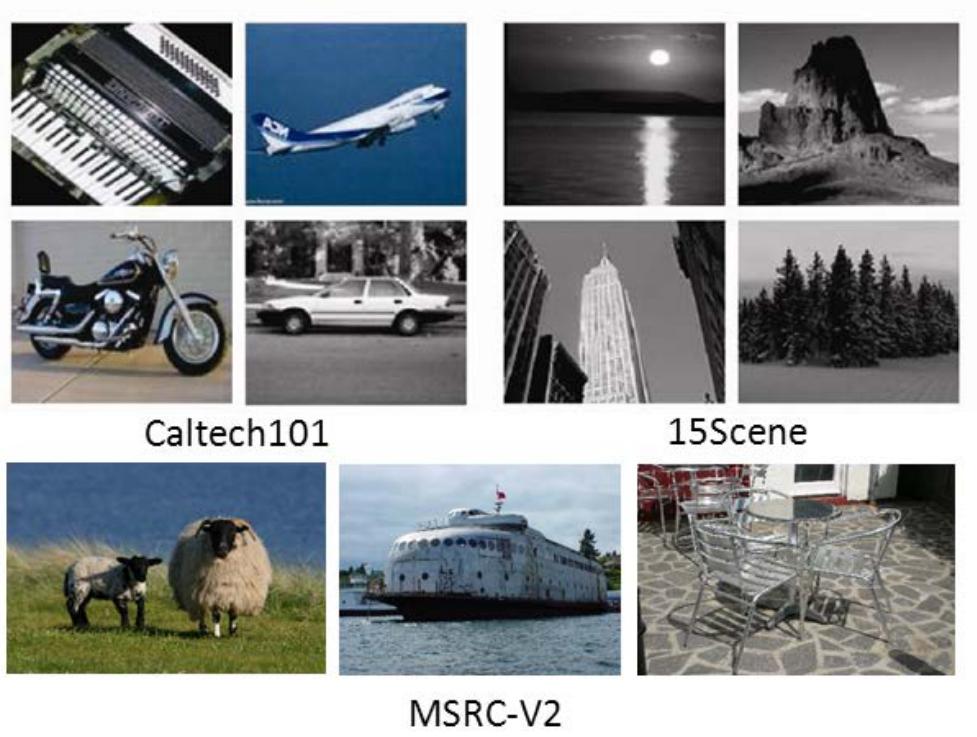

Figure 3.3: Some example images from the Caltech101, 15Scene and MSRC-v2 image data sets.

3.4.1 Image data sets

For this work, we use MSRC-v2, Caltech101 and 15 Scene data sets for experiments. Figure 3.3 shows some example images from these data sets.

This subsection provides short descriptions of these image data sets.

MSRC-v2: In this data set, there are 591 images that accommodate 23 different categories. All the categories in the images are manually segmented. Different subsets of these categories have been used by several authors to derive a classification problem [75, 80].

15Scene: This data set [50, 53, 68] comprises indoor (i.e. office, kitchen, bedroom etc.) and outdoor (i.e. beach, mountain, tall building etc.) scenes. Images were collected from different sources predominantly from Internet, COREL collection and personal photographs. Each category has 200 to 400 images, and the average image size is 300×250 pixels.

Caltech101: The Caltech101 data set [52] has 102 object classes. It is one of the most diverse object database available. However, this data set has some limitations. Namely, most images feature relatively little clutter and possess homogeneous backgrounds. In addition, there are very less variations among the objects of the same category. Despite the limitations,

this data set is quite a good resource containing a lot of interclass variability.

3.4.2 Implementation Details

For MSRC-v2 data set we use a 15 category problem as used in [55, 75]. We use a filter-bank responses for feature extraction as in [55, 75]. The training and testing sets are also chosen in accordance with those works for the sake of comparison. For the other data sets, we follow the experimental setup consistent with [50]. Thus, we use single scale dense detector and SIFT descriptor for feature extraction. To be able to compare our results with other spatial representations, we use the standard BoVW representation (hard assignment). Thus, for all the data sets, we apply K-means on the descriptors to construct the vocabularies. Each descriptor is then mapped to the nearest visual word based on euclidean distance. Support Vector Machine (SVM) is used to perform the classification tasks. We use the intersection kernel [81] and the one-vs-all rule where multi-class classification is necessary. The cost parameter C was optimized for all the experiments using a 10-fold method on the training set. Note that, this representation does not require any quantization for 2nd order descriptors as opposed to [75]. So, the output of our algorithm is directly fed into the classification algorithm.

3.4.3 Parameter tuning

In our approach, three parameters(M_θ, M_R and σ) have to be set to compute classification results. We study their influence on this section. In figure 3.4, on the left, we plot the effect of the number of angle bins(M_θ) and distance bins(M_R) on classification accuracy on 15-scene and Caltech101 data sets for a vocabulary size of 200. A 36 bins (9×4) spatial histogram appears to be a good compromise for both datasets. Considering finer quantization does not improve the accuracy significantly, but highly increases feature dimension. On the right of figure 3.4, we show the effect of the weighting parameter σ on accuracy. For a very low σ, not all similar patches are taken into account and for a higher σ, there are patches which may not be similar and could be regarded as noise. Whatever the data set, the value σ=0.3 gives the best results and will be used in the following sections. This value is related to the descriptor in use (SIFT in this case).

3.5 Results

In this section, we first study the performance of our SPS_{ad} image representation and its HPS_{ad} and SPS_{ad}+ derivatives on the Caltech101 and

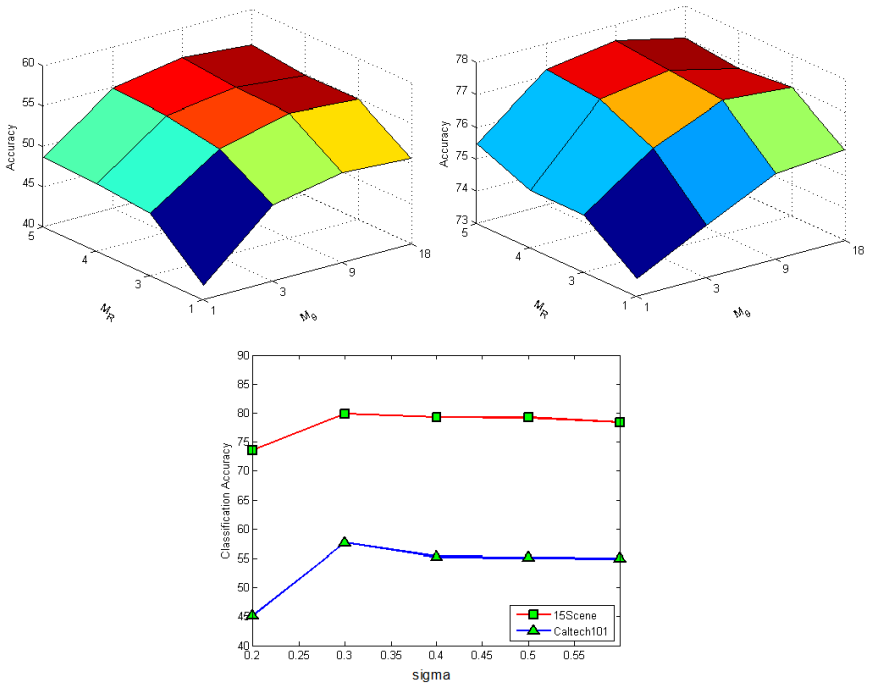

Figure 3.4: Parameter tuning for SPS_{ad} representation. On the top, the influence of number of bins for Caltech101(left) and 15Scene(right) data sets and at the bottom, the influence of σ for the same data sets.

15Scene data sets. Next, we compare SPS_{ad} with other similar spatial representations on the MSRC-2 dataset.

3.5.1 Performance evaluation of SPS_{ad} representation

Here, we first analyze the performance for SPS_{ad} representation on Caltech101 and 15Scene data sets(The MSRC-2 dataset is used in section 3.5.2 to compare with [55, 75]). We show the classification performance gain for SPS_{ad} over BoVW and HPS_{ad} representation and discuss the results. Next, we compare SPS_{ad}+ with SPR and some other spatial descriptors obtained from combination with SPR.

Table 3.1 shows results on Caltech101 and 15 Scene data sets for 3 different vocabulary sizes. From these results, it is clear that for each data set the SPS_{ad} representation improves the results over BoVW and HPS_{ad} representation for all the vocabulary sizes. For larger dictionaries, spatial information does not seem to be as effective as in the smaller ones. Indeed, comparing SPS_{ad} over BoVW, for Caltech101 and 15Scene, gains are respectively about 14% and 7% for a vocabulary size of 100 and decrease to 10.25% and 3.3% for a vocabulary size of 1000. This was also observed in some of the previous works [50, 97, 105].

Dataset	Voc. Size	BoVW		HPS_{ad}		SPS_{ad}	
		μ	σ	μ	σ	μ	σ
Caltech101	100	39.83%	1.32	53.01%	1.1	53.91%	1.23
	200	41.12%	1.06	55.3%	0.9	57.47%	1.00
	400	45.56 %	1.54	52.11 %	1.38	57.62 %	1.38
	1000	48.08 %	1.42	51.28 %	1.58	58.33 %	1.41
15 Scene Dataset	100	70.83%	0.6	76.11%	0.46	77.96%	0.46
	200	72.2%	0.6	77.52%	0.59	79.38%	0.67
	400	75.7 %	0.33	78.11 %	0.5	79.58 %	0.8
	1000	76.82 %	0.61	77.91 %	0.7	80.11 %	0.56

Table 3.1: Classification accuracy comparison among BoVW representation, HPS_{ad} and SPS_{ad}. Mean (μ) and Standard Deviation (σ) over 10 individual runs are presented.

It is interesting to note that with the increase of vocabulary size both BoVW and SPS_{ad} representation increase even though the gain of SPS_{ad} over BoVW gets smaller. However, HPS_{ad} reaches an optimal and decreases with the increasing vocabulary size. The reason is, for larger dictionaries intratype words become scarce(one cluster is divided into multiple clusters) and thus HPS_{ad} cannot provide important spatial information. On the other hand, SPS_{ad} should always be able to add spatial information into the BoVW representation regardless of the state of the vocabulary.

Now, we compare the combination descriptor $SPS_{ad}+$ with SPR and different other approaches [99, 102] that also propose combination descriptors based on SPR. Along with fundamental similarities with $SPS_{ad}+$, these methods also use similar setup to us namely, dense sampling as detector, SIFT as descriptor, K-means for vocabulary construction, same number of visual words and histogram based representation, thus facilitates fair comparison. Table 3.2 shows the comparison among all the mentioned methods for Caltech101 and 15 Scene datasets. We can see that the global distribution of visual words is complementary to the global correspondence and our method outperforms SPR and the other methods in all cases. Interestingly, the compact version $SPS_{ad}^{1800}+$ also outperforms all the other existing methods under similar setup with a lower feature dimensionality than these methods.

3.5.2 Comparison between SPS_{ad} and other spatial methods

Here, we compare our method with Savarese et al. [75] and Liu et al. [55]. These two works are the most notable among those which concern modeling spatial relationships among the visual words. They rely on the use of new features composed of pairs (or higher number) of words having a specific relative position in order to build spatial histograms. Note that, contrary to

Methods	Caltech101	15 Scene Dataset	feature dimensionality
SPR Single Level (L=2)	63.4%*	79.4%*	3200
SPR Entire Pyramid (L=2)	64.6%*	81.1%*	4200
SPS_{ad}+	**68.1%**	**83.7%**	11400
SPS_{ad}^{1800}+	**67.5%**	**83%**	1800
$PIWAH$+ [40]	67.1%	82.5%	5000
Zhang et al. [102]	65.93%	81.5%	9600
Yang et al. [99]	X	82.5%	Not Clear
Spatial Fisher Vector [47]	X	82.0%	268800

Table 3.2: Classification accuracy(%) comparison among SPR , SPS_{ad}+ and two other methods for Caltech101 and 15 scene dataset. Results with * are taken from [50]. a 'X' means that the result is not present in the corresponding work.

Criteria of Comparison	SPS_{ad}	Savarese et al. [75]	Liu et al. [55]
Accuracy	83.5%	81.1%	83.1%
Global Spatial Association	Y	N	N
2nd Order Feature Quantization	N	Y	N
Pre-processing/Feature Selection Step	N	Y	Y

Table 3.3: Comparison among existing methods on a 15 class problem derived from MSRC-V2 dataset.

our method, the previous approaches do not directly incorporate the spatial information of pair of similar words. We focus on several criteria to compare our work with the mentioned ones. The table 3.3 shows the details of the comparisons on MSRC-v2 dataset for 400 visual words. For this dataset, SPS_{ad} representation provides the best classification results. Our method also holds different other advantages over the existing methods. For example, Liu et al. [55] integrates feature selection and spatial information extraction to boost recognition rate. However, as the spatial feature extraction becomes a part of the learning step, the modification in the training set would lead to recomputation of features and thus making it difficult to generalize. Let's also note that, SPS_{ad} models only global association and unlike Savarese et al. [75], does not require a 2^{nd}-order feature quantization. As the previous approaches fail to incorporate the spatial information of similar pairs properly, our approach is complementary to these approaches as well.

3.6 Conclusion

In this chapter, we proposed a new method to model global spatial distribution of visual words and improved the standard BoVW representation. This method exploits spatial orientations and distances of all pairs of similar descriptors in the image. The evaluation was made on an image classifica-

tion task, using an extensive set of standard data sets. Experiments confirm that our model outperforms the standard BoVW approach and the existing spatial methods on all the data sets. Compared to the global correspondence methods as SPR, our model brings complementary information. In this case, we outperform all of the methods that do the same. One interesting future direction could be to extend our method to advanced BoVW encoding techniques [73, 98]. Spatial information provided by multiple cues e.g. color and shape, is also promising as a future direction.

Chapter 4

Discriminative Color Descriptors

Résumé: La description couleur représente un réel challenge à cause de la variabilité importante des valeurs RVB fortement influencées par les ombrages, la spécularité, les changements de couleur de l'éclairage ou encore les changements de point de vue. Traditionnellement, cette tche est abordée en partant de modèles physiques et en en déduisant des invariants. L'inconvénient d'une telle approche est que des couleurs initialement distinguables dans l'espace couleur sont projetées en un même point de l'espace photométrique invariant et ne peuvent plus être distinguées. Ce résultat conduit à une chute du pouvoir discriminant de la description couleur. Dans ce chapitre, nous présentons une approche de description couleur basée sur la théorie de l'information. Nous regroupons les valeurs de couleurs en exploitant leur pouvoir discriminant étant donné un problème de classification. Ce regroupement (clustering) a pour objectif explicite de minimiser la chute d'information mutuelle de la représentation finale. Nous montrons qu'une telle description couleur permet d'apprendre automatiquement un certain degré d'invariance photométrique. Nous montrons également qu'une représentation couleur universelle, établie à partir d'autres bases de données que celle traitée, permet d'aboutir à des résultats compétitifs par rapport à l'état de l'art. Les résultats expérimentaux démontrent que le descripteur couleur proposé obtient des performances supérieures à celles obtenues par des invariants photométriques de l'état de l'art. De plus, il est établi que combiné avec des descripteurs de forme, notre descripteur couleur fournit d'excellents résultats sur quatre bases de données bien connues, PASCAL VOC 2007, Flowers-102, Stanford dogs-120 and Birds-200.

Abstract: Color description is a challenging task because of large variations in RGB values which occur due to scene accidental events, such as shadows, shading, specularities, illuminant color changes, and changes in

viewing geometry. Traditionally, this challenge has been addressed by capturing the variations in physics-based models, and deriving invariants for the undesired variations. The drawback of this approach is that sets of distinguishable colors in the original color space are mapped to the same value in the photometric invariant space. This results in a drop of discriminative power of the color description. In this chapter we take an information theoretic approach to color description. We cluster color values together based on their discriminative power in a classification problem. The clustering has the explicit objective to minimize the drop of mutual information of the final representation. We show that such a color description automatically learns a certain degree of photometric invariance. We also show that a universal color representation, which is based on other data sets than the one at hand, can obtain competing performance. Experiments show that the proposed descriptor outperforms existing photometric invariants. Furthermore, we show that combined with shape description these color descriptors obtain excellent results on four challenging data sets, namely, PASCAL VOC 2007, Flowers-102, Stanford dogs-120 and Birds-200. [1]

4.1 Introduction

The description of color is an important problem for a wide range of computer vision applications. In many of these applications the BoVW image representation is used. In such representations, color next to shape, was found to be an important cue [39, 86]. In this chapter, we propose a new method to learn discriminative color descriptors.

Color description is difficult due to the many scene accidental events which influence its measurement. These events include shadows, illuminant changes, variations in scene geometry and viewpoint, and acquisition device specifications. This has sparked an extensive literature on photometric invariance which aims to describe color invariants with respect to some of these variations [27]. Based on reflection models [76] or assumptions on the illumination [21], invariance with respect to shadow, shading, specularities and illuminant color can be obtained. However, photometric invariance is gained at the cost of discriminative power. Therefore, in designing color representations, it is important to weight the gains of photometric invariance against the loss in discriminative power.

We propose to learn color descriptors which have optimal discriminative power for a specific classification problem. The problem of learning a color descriptor is equal to finding a partition of the color space. Our approach relies on the Divisive Information-Theoretic Clustering (DITC) algorithm proposed by Dhillon *et al.*[18] to learn this partition. We adapt

1. The content of this chapter appeared in the proceeding of the Computer Vision and Pattern Recognition(CVPR), 2013 [42].

this algorithm to ensure that the final clusters are smooth and connected. Considering all the values in the L*a*b* cube, we aim to join values in this L*a*b* cube driven by the discriminative power of the final representation, the latter being measured using information theory. We distinguish two variations. Firstly, the specific color descriptor which is optimized for a single data set. Secondly, a universal color descriptor which is trained on multiple data sets, thereby representing a wide range of real-world data sets. The advantage of universality is that users can run the learned mapping for an unknown data set without the effort of learning a data set specific color representation. In experimental results we will show that these discriminative color descriptors outperform purely photometric color descriptors, and that combined with shape description they can obtain state of the art results on several data sets.

4.2 Background and motivations

In this section, we discuss the background and motivation of this work. We start with a review of photometric invariance based color descriptors. Next, we present the classical problem of invariance vs discriminative power in case of color from an information theoretic point of view. Finally, we present the color names descriptor - which is the state-of-the-art color descriptor. We conclude this section summarizing our motivations.

4.2.1 Photometric invariance based color descriptors

Representing color information is very challenging as it varies largely depending on the viewing conditions (illumination, geometry etc.). The study of photometrically invariant color descriptors starts from color formation models. Formation of color depends on three different parameters, namely, illuminant, surface reflectance and spectral sensitivities of sensor. Let $s(\mathbf{x}, \lambda)$ denote the spectral power distribution (SPD) of the light source that illuminates a surface at a spatial position $\mathbf{x}$ which has reflectance $r(\mathbf{x}, \lambda)$. If the camera is having spectral sensitivities $c_n(\lambda)$ ($n = R, G, B$ for a tri-chromatic system), its responses are each obtained by integrating over the visible spectrum:

$$\rho_n = m^b(\mathbf{x}) \left(\int_{vis} r(\mathbf{x}, \lambda) s(\mathbf{x}, \lambda) c_n(\lambda) \mathrm{d}\lambda \right) \tag{4.1}$$

where m^b is geometric dependence of the reflectance properties. Equation 4.1 is commonly known as a Lambertian model [49]. This model predicts that the pixel values for a single colored object lie on a line passing through the origin of the RGB cube. This assumption holds to some extent in case of matte surfaces (e.g. chalk, paper). However, for material surfaces with specularities a.k.a glossy surfaces, it does not hold. To this end, Shafer proposed

the dichromatic reflection model(DRM) [76] which is a better assumption for matte and glossy color formation:

$$\rho_n = m^b(\mathbf{x}) \left(\int_{vis} r(\mathbf{x}, \lambda) s(\mathbf{x}, \lambda) c_n(\lambda) \mathrm{d}\lambda \right) + m^i(\mathbf{x}) \left(\int_{vis} s(\mathbf{x}, \lambda) c_n(\lambda) \mathrm{d}\lambda \right) \tag{4.2}$$

DRM dictates that, the sum of a diffused and a specular component forms the actual color. In Equation 4.2 $m^b(\mathbf{x})$ and $m^i(\mathbf{x})$ denotes the contribution of the diffused and the specular component respectively. These parameters depend on the viewing angle, direction of incident illuminant and surface orientation. Note that, if $m_i = 0$ in Equation 4.2, we obtain Equation 4.1.
Color features are based on these color formation models. Next, we are going to describe some color descriptors based on the understanding of color formation.

4.2.1.1 rg-histogram

It could be shown from equation 4.1 that, $r = \frac{\rho_R}{\rho_R+\rho_G+\rho_B}$ and $g = \frac{\rho_G}{\rho_R+\rho_G+\rho_B}$ become invariant to change in illumination intensity and thus in the variation of shadow and shading. The rg-components provide some invariance under the lambertian assumption(equation 4.1). Uniformly quantized histogram of such a normalized image could be made and then used as an intensity invariant color descriptor.

4.2.1.2 Hue-histogram

Under the assumption of DRM, it could be shown that hue is invariant to surface orientation, illumination direction and illumination intensity (although not to illumination color). However, in the HSV color space, Van de Weijer et al. [93] have shown that the certainty of the hue is inversely proportional to the saturation. Therefore, the hue histogram is made more robust by weighing each sample of the hue by its saturation. There exist multiple versions of hue-histogram descriptors in the literature [86, 92].

4.2.2 Photometric invariance versus discriminative power

Color feature design has been mainly motivated from photometric invariance perspective [24, 26, 92]. It is based on the observation that colors in the world are dependent on scene incidental events such as scene geometry, varying illumination, shadows, and specularities. To obtain invariance with respect to these effects, photometric invariant features can be derived.

But one could wonder what the cost of photometric invariance is. Mapping multiple RGB values to the same photometric invariance will potentially lead to a drop in discriminative power. This aspect of photometric

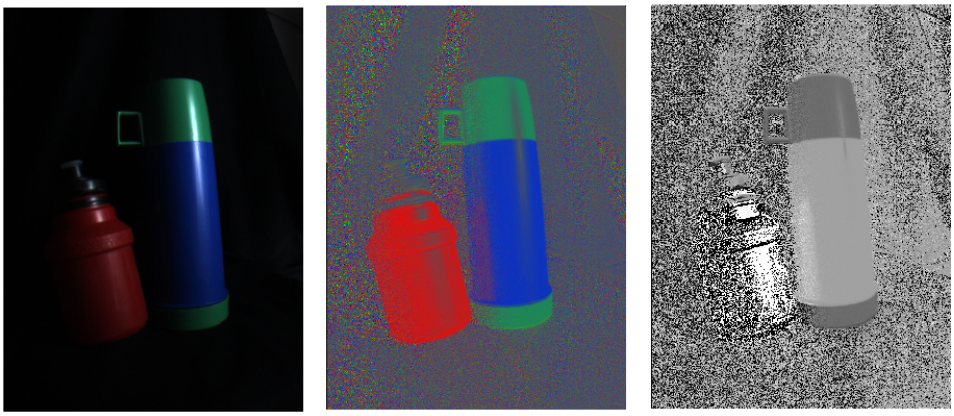

Figure 4.1: Invariance of rg-normalized image and hue. The rg-normalized image is invariant to shadow and shading but not to specularities. Hue is comparatively more invariant to specularities. Note the uncertainity in the hue when saturation is low(achromatic colors on the objects and the background).

invariance has received relatively little attention. Stability and noise sensitivity were measured by Stokman et al. [28]. Geusebroek et al. [26] showed that with increasing invariances fewer Munsell patches [2] could be distinguished. Here we will analyze the drop in discriminative power in a more principled way by means of information theory.

We discretize our initial color space into m color words $W = \{w_1, ..., w_m\}$. In our case m is equal to $m = 10 \times 20 \times 20 = 4000$ of equally spaced grid points in the L*a*b* cube. Consider we have a data set with l classes $C = \{c_1, ..., c_l\}$. These classes are represented by histograms over the color words. The discriminative power of the color words W on the problem of distinguishing the classes C can be computed by the mutual information:

$$I(C, W) = \sum_i \sum_t p(c_i, w_t) \log \frac{p(c_i, w_t)}{p(c_i)\, p(w_t)} \tag{4.3}$$

where the joint distribution $p(c_i, w_t)$ and the priors $p(c_i)$ and $p(w_t)$ can be measured empirically from the data set.

The mutual information measures the information that the words W contain about the classes C. Now consider we divide the words W into k clusters $W^C = \{W_1, ..., W_k\}$ which are invariant with respect to some physical variation. Each cluster W_j represents a set of words. Then Dhillon

2. The Munsell color system is a color space that specifies colors based on three color dimensions: hue, value (lightness), and chroma. It was created by Professor Albert H. Munsell.

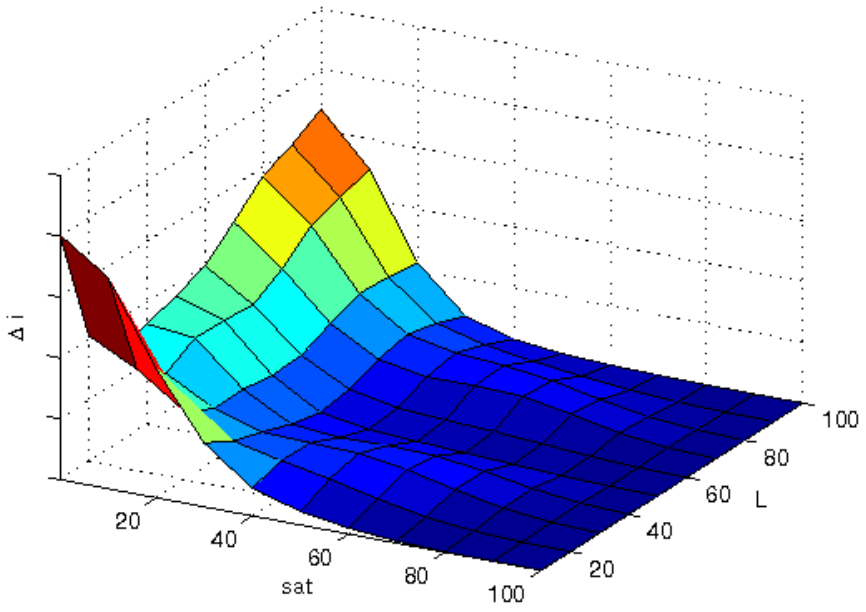

Figure 4.2: Graph showing the drop in mutual information for the flower data set caused by grouping bins with equal chromatic values (a and b). From the graph it can be seen that the drop of mutual information is largest for low saturated points, especially with low and high lightness (L).

et al.[18] proved that the drop of mutual information caused by clustering a word w_t to cluster W_j (in our case based on photometric invariance) is equal to:

$$\Delta i = \pi_t KL\left(p\left(C|w_t\right), p\left(C|W_j\right)\right) \tag{4.4}$$

where the Kullback-Leibler (KL) divergence is given by:

$$KL\left(p_1, p_2\right) = \sum_{x \in X} p_1\left(x\right) \log \frac{p_1\left(x\right)}{p_2\left(x\right)} \tag{4.5}$$

and $\pi_t = p\left(w_t\right)$ is the word prior.

The above Equation 4.4 provides a way to assess for each color value the drop in discriminative power Δi which is caused by imposing photometric invariance. In Figure 4.2 we plot the drop in mutual information which occurs when we look at a photometric invariant representation with respect to luminance. This is simply obtained by defining clusters as the set of bins of equal (a, b) values, computing the $p(C|W_j)$ of each cluster, and computing Δi with Equation 4.4. We plot the drop in mutual information as a function of lightness L and saturation $sat = \sqrt{(a^2 + b^2)}$. The plot is based

on the Flower data set [65] but similar results were observed for other data sets. The plot tells a clear story: the largest loss of discriminative power is occurring for achromatic (or low saturated) colors as is clear from the ridge at $sat = 0$. Even though these achromatic colors cannot be distinguished from a photometric invariance point of view (since they can be generated from each other by viewpoint or shadow variations), this analysis shows that they contain discriminative power. This leads us to investigate an alternative approach to color feature computations based on discriminative power presented in section 4.3. Some aspects of the proposed descriptor is similar to the color names descriptor [87]. So, in the next subsection, we discuss the color names descriptors, which is the state-of-the-art color descriptor.

4.2.3 The color names descriptor

We have discussed the color names descriptor in brief in section 2.2.2.2. Color names are linguistic labels humans use to communicate the colors in the world. Examples of color names are 'red', 'black', 'turquoise' etc. Van de Weijer et al. [87] have proposed a method to automatically learn the eleven basic color names of the English language from Google images. The color names descriptor [87] CN is defined as a vector containing the probability of a color name given an image region $\mathcal{R}$.

$$CN = \{p\left(cn_1 \mid \mathcal{R}\right), p\left(cn_2 \mid \mathcal{R}\right), \ldots, p\left(cn_{11} \mid \mathcal{R}\right)\} \tag{4.6}$$

with

$$p\left(cn_i \mid \mathcal{R}\right) = \frac{1}{P} \sum_{x \in \mathcal{R}} p\left(cn_i \mid f\left(x\right)\right) \tag{4.7}$$

where cn_i is the i-th color name, x are the spatial coordinates of the P pixels in region $\mathcal{R}$, $f(x)$ is the color values (e.g. $\{L^*, a^*, b^*\}$) at the position x and $p\left(cn_i \mid f\right)$ is the probability of a color name given a pixel value. The probabilities $p\left(cn_i \mid f\right)$ are computed from a set of images collected from Google. To learn color names, 100 images per color name are used. To counter the problem of noisy retrieved images, PLSA approach is used by [87]. To describe a local patch with this descriptor, the average of all the pixels are computed and then represented with a 11D color names vector(Figure 4.3).

In the original work [87], the authors use a soft assignment of color names such that each tri-chromatic color is assigned to 11 probability values. Each value is the probability of a particular color name to be associated with that tri-chromatic color. However, one can only consider the highest probability component such that each tri-chromatic color is associated with a single color name. This way a partition of the color space into eleven regions could be obtained. Then, an eleven dimensions local color descriptor can be deduced simply by counting the occurrence of each color name

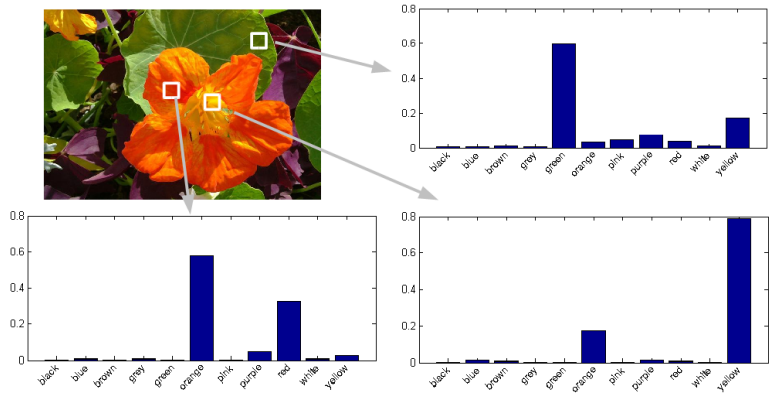

Figure 4.3: Three examples of color name descriptors calculated from 3 local patches of an image from the Flower102 data set. The highest probability color name very often dominates the distribution.

over a local neighborhood. Analyzing the clusters of RGB values which are appointed to a color name (let us consider 'red' for example), we note that these clusters possess a certain amount of photometric invariance. Multiple shades of red are all mapped to the same color name 'red'. For example, if we place a 'red' surface under different natural light sources with different color temperatures, all the variations of that surface color are expected to be members of the 'red' cluster. However, when moving towards darker 'reds', at a certain point the values will be mapped to the color name 'black' instead, and the photometric invariance breaks down. Recently, color names were found to compare favorably against photometric invariant descriptions on several computer vision applications, such as image classification [39] and object detection [36]. These results show that focus on photometric invariance which is at the basis of many color descriptors might not be optimal. They further suggest that discarding discriminative power of the color representation will deteriorate final results. In Figure 4.4 three such clusters for 'red', 'white' and 'yellow' colors are shown.

It is intuitive that, learning additional color names would give more discriminative power to the descriptor. However, color linguists and psychologists have argued that after the 11 basic color names, there is no known ordering to extend from these set of colors. This is why, increasing the number of colors from 11, in the color names descriptors is an open problem.

4.2.4 Remarks and conclusion

The primary motivation of our work is driven by the observation that the existing color descriptors are fundamentally designed for invariance and not optimized to be discriminative. Color descriptors could be obtained

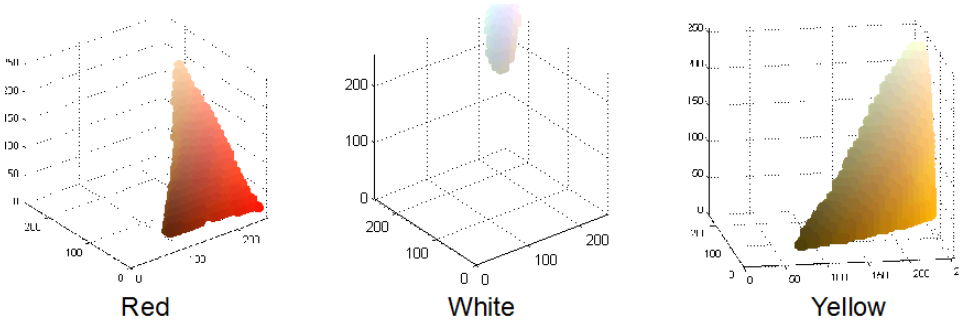

Figure 4.4: Red, white and yellow color clusters are shown. These clusters were obtained from the original work [87]. Note the compactness and smoothness of the clusters which are essential to obtain photometric invariance.

in the form of clusters in the color space as in the color names descriptor. However, just like the photometrically invariant descriptors, the color names descriptor is also not optimized for discriminative power. So, we set our goal to compute clusters in the color space, which are discriminative given a set of labeled training images. To achieve this goal, in the next section we outline our approach of discriminative color feature computation, which clusters color values together based on discriminative power on a training data set. The expectation is that discriminative clustering will automatically lead to a certain amount of photometric invariance by clustering values of similar hue together. However, as explained in section 4.2.2, clustering similar hues together results into significant drop in discriminitive power around the achromatic axis. So, in that region, we expect additional clusters to arise using our approach, to reduce the drop in discriminative power caused by the clustering.

4.3 Discriminative color representations

In this section we discuss our discriminative approach to color representation learning. We first explain divisive information-theoretic feature clustering (DITC) proposed by Dhillon et al.[18]. Next, we adapt the algorithm to find connected clusters in L*a*b* space.

4.3.1 The DITC algorithm

The DITC algorithm provides a way to cluster features into a smaller set of clusters, where each cluster contains a number of features from the original set. The clustering is performed in such a way as to minimize the decrease of mutual information (Equation 4.3) of the new and more compact

representation. The total drop of mutual information caused by clustering the words, using Equation 4.4, is equal to

$$\Delta I = \sum_{j} \sum_{w_t \in W_j} \pi_t KL\left(p\left(C|w_t\right), p\left(C|W_j\right)\right). \tag{4.8}$$

Hence the clusters W_j which we seek are those which minimize the KL divergence between all words and their assigned cluster (weighted by the word prior). In our case the words represent L*a*b* bins of the color histogram. This color space is used because of its perceptual uniformity. Minimizing Equation 4.8 is equal to joining bins from the L*a*b* histogram in such a way as to minimize the ΔI. L*a*b* bins which have similar $p\left(C|w_t\right)$ are joined together.

An EM like algorithm is used to optimize the objective function 4.8. The algorithm alternates between two steps.

1. Compute the cluster means with

$$p\left(C|W_j\right) = \sum_{w_t \in W_j} \frac{\pi_t}{\sum\limits_{w_t \in W} \pi_t} p\left(C|w_t\right). \tag{4.9}$$

2. Assign each word to the nearest cluster according to

$$\mathbf{w}_t^* = \arg\min_j KL\left(p\left(C|w_t\right), p\left(C|W_j\right)\right). \tag{4.10}$$

The new cluster index for word w_t is given by $\mathbf{w}_t^*$.

The algorithm is repeated until convergence. For more details we refer to [18].

The DITC algorithm has been studied in the context of joining color and shape features into so-called Portmanteau Vocabularies by Khan et al. [37]. In this work, we use the DITC algorithm for a different purpose, namely to automatically learn discriminative color features. In addition, we propose two adaptations to the DITC algorithm.

4.3.2 Learning compact color representations

The original DITC clustering algorithm does not take into account the position in the L*a*b* space of the words. As a consequence, the algorithm can join non-connected bins. It is known that photometric variations result in connected trajectories [88]. Therefore when learning photometric invariants we expect them to be connected. In addition, connectivity has several conceptual advantages: it allows for comparison to photometric invariance, comparison with color names (CN), semantic interpretation (human color names are connected in the L*a*b* space), and comparison with human perception (e.g. MacAdam Ellipses). Therefore we propose to adapt the DITC

algorithm to ensure that the clusters are connected in the L*a*b* space. As a second adaptation we enforce smoothness of the clusters which prevents them from over fitting to the data. Both objectives can be translated into an additional energy term which can be added to the objective function of Equation 4.8.

Let $\mathbf{w}_t$ be the cluster number assigned to word w_t, and $W_{\mathbf{w}_t}$ is the cluster to which w_t is assigned, then the cost of choosing a certain cluster assignment according to Equation4.8 is equal to

$$\psi_t^I(\mathbf{w}_t) = \pi_t KL\left(p\left(C|w_t\right), p\left(C|W_{\mathbf{w}_t}\right)\right). \tag{4.11}$$

In this standard objective function, the relation of the words is not taken into account, and the final clusters W^C can — and most likely will — contain words which are not connected in color space. We enforce connectivity by introducing a cost for not being connected to the principal component of the cluster. The principal component $\mathcal{P}_j$ of a cluster W_j is defined as the connected component with the highest prior mass (the component for which the sum of the priors of its words is largest). Words which are not connected to the principal component of the cluster will have an additional cost for taking on this cluster assignment. We identify words connected to the principal component by $\mathcal{P}'_j$ and they are computed with a morphological dilation with a 26-connected structuring element b:

$$\mathcal{P}'_j = \mathcal{P}_j \oplus b. \tag{4.12}$$

This type of dilation is justified because we use equi-quantized bins on a uniform L*a*b* color space. After this dilation $\mathcal{P}'_j$ contains all words connected to the principal component of cluster j. We add a penalty term to all the color bins which are not part of $\mathcal{P}'_j$ according to

$$\begin{gathered}\psi_t^C(\mathbf{w}_t) = \alpha_C \cdot \left(1 - f^t(\mathbf{w}_t)\right) \\ \text{Where} \qquad f^t(\mathbf{w}_t) = 1 \qquad \text{if} \qquad w_t \subset \mathcal{P}'_{\mathbf{w}_t}\end{gathered} \tag{4.13}$$

With a sufficiently high choice of the constant α_C , this energy will eliminate non-connected assignments, and result in a final clustering of the features into connected clusters. We present a toy example to understand this step in Figure 4.5.

To enforce our second objective of smoothness of the color representation we introduce a pairwise cost according to

$$\psi(\mathbf{w}_s, \mathbf{w}_t) = \begin{cases} 0 & \text{if } \mathbf{w}_s = \mathbf{w}_t \\ \alpha_D & \text{otherwise} \end{cases} \tag{4.14}$$

Now consider a certain labeling for all words $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_m\}$ then the cost of this labeling can be written to be

$$E(\mathbf{w}) = \sum_t \left(\psi_t^I(\mathbf{w}_t) + \psi_t^C(\mathbf{w}_t)\right) + \sum_{(s,t)\in\varepsilon} \psi(\mathbf{w}_s, \mathbf{w}_t) \tag{4.15}$$

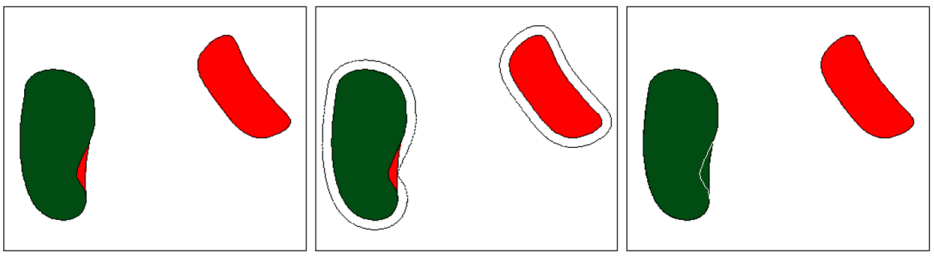

Figure 4.5: A 2-cluster toy example in 2D to demonstrate the working principle of the dilation step of our algorithm. The clusters are color coded i.e. the red and green regions are different clusters. The left image shows the previous state of the clusters, note the red part in the vicinity of the green cluster. The middle image shows the dilation step, where the principal components are dilated and for each cluster a penalty term is added to all the parts not inside the dilated region. The right image shows the current state of the clusters. Now the non-connected part of the red cluster is a part of the green cluster (the white border is used for illustration purpose only).

where ε is the set of all connected words s and t.

The two step algorithm(Equation 4.9 and Equation4.10) has to be slightly adapted to minimize this objective function. Step one remains unchanged and computes the cluster means. In step two, we aim to find $\mathbf{w}^*$ which minimizes Equation 4.15. This can be done with a graph cut algorithm where the nodes are the words (or bins of L*a*b* histogram) and the vertices connect neighboring nodes. After the optimal assignment $\mathbf{w}^*$ is found, the algorithm returns to step one until convergence. Like the original DITC algorithm, the most expensive step in our proposed version is also the first step (Equation 4.9) of the algorithm. Thus, the time complexity of the modified version is $O(mlk\tau)$, where m is the number of words, l is the number of classes, k is the number of desired clusters and τ is number of iterations needed to reach convergence.

4.3.3 Convergence

Our optimization of the objective function of Equation4.15 is obtained by iteratively applying the two steps above. However, when we dilate all clusters (to define the connected bins), it could theoretically happen, that for some bins which change label, the bin to which they were connected also changes label. This could lead to unconnected components, and would activate the cost defined in Equation4.13, and lead to an increasing objective function. This could be addressed by changing labels one bin at a time, but this would be computationally very costly. Practically, we run the iterations until no change in the labeling occurs. For the three data sets (and

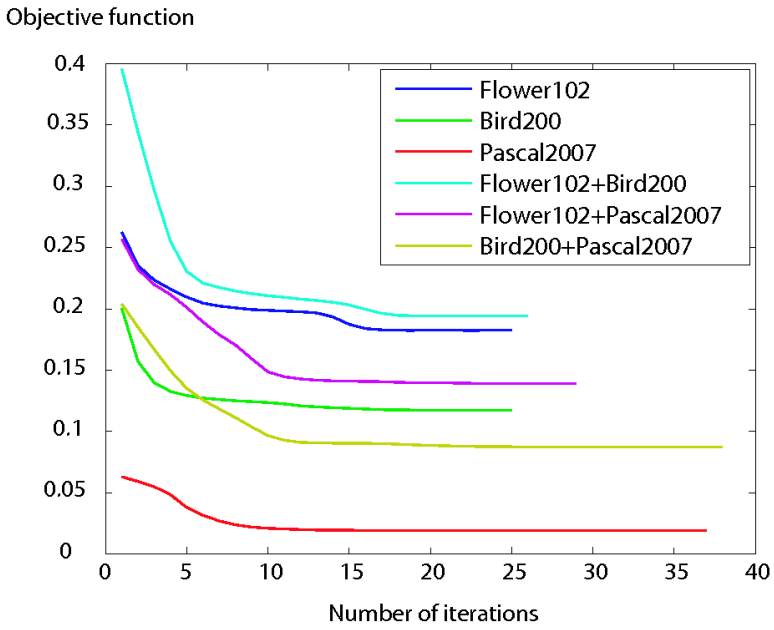

Figure 4.6: Evolution of the objective functions for some image sets until convergence.

their three combinations) used in this work, we verified that the final color descriptors were connected. Figure 4.6 shows the evolution of the objective function for the six runs until convergence.

4.3.4 Photometric invariance of learned clusters

Instead of imposing photometric invariance, as is generally done, we follow an information theoretic approach which maximizes the discriminative power of the final representation. The underlying idea being that clustering color bins based on their discriminative power would automatically learn a certain degree of photometric invariance. Here, we verify that this has happened by analyzing the cluster assignments for two images.

We learn a 11-dimensional discriminative color descriptor for the Flower data set. Next, we apply the descriptor on two images of the data set. The results are depicted in Figure 4.7. Here, we replace the color of each pixel by the average color of all the pixels assigned to the same cluster. We can see that clusters are constructed so that they allow to discriminate flowers from background and leaves while providing some robustness across some photometric variations. For example, note that the pixels under the shadows caused by the wrinkles on the yellow petals are assigned to the same cluster and the stamen part of the red flower is mapped to one cluster in spite of the photometric variations in the pixels. Also, the dark pixels that introduce

Figure 4.7: Examples of cluster assignment on two images from the Flower data set.

most noises into photometric invariance representation are assigned to a separate cluster. The photometric invariance can also be observed from the bottom row of Fig. 4.8 where we see that pixels with similar *hue* but varying intensity are grouped together.

4.4 Universal color descriptors

In a seminal work named 'Basic color terms: their universality and evolution' the linguists Berlin and Kay [7] show the universality of the human basic color names. With universality they refer to the fact that the basic color names which are used in different cultures have a similar partition of the color space: the Arab azraq refers to a similar set of colors as the English blue. In the context of descriptors, we will use the term universality to refer to descriptors which are not specific to a single data set. Universality is one of the more attractive properties of the computational color names [6, 87]. As a consequence of universality, users are not required to learn a new color representation for ever new data set and can just apply the universal color representation to their problem.

In the previous section, we showed how to learn discriminative color features. Applying the above algorithm to a specific data set results in a color representation which is data set specific in the sense that it is optimized to discriminate between the classes of that data set. The same setup can be used to learn universal color vocabulary by joining several training sets together to represent the real-world. We learn such a description combining

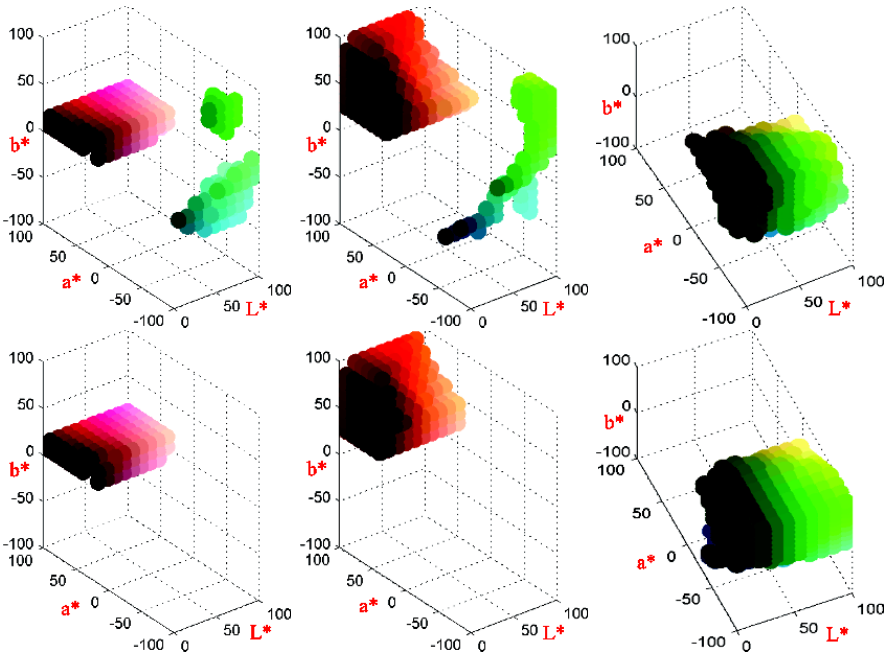

Figure 4.8: The clusters of the first and second row are computed from the Flower102 training set, by the original DITC algorithm and the proposed method respectively. Note the compactness and smoothness of the color clusters computed by the proposed method.

the training sets of Flower102, Bird200 and PASCAL 2007 data sets. An advantage over the existing computational color names [87] is that we are not limited to eleven color names and can freely choose the desired dimensionality. We make the universal color descriptors available for the settings with 11, 25, and 50 clusters.

In the experiments we will investigate universal color descriptors, and compare them to specific color descriptors. We will do so by training the universal color descriptor from other data sets than the one currently considered. Universality is expected to result in a drop of performance since the descriptor cannot adapt to the specificity of the data set. However, if the drop is small the advantages of a universal representation can outweigh the drop in performance.

4.5 Experimental results

In the next few subsections, we discuss experimental details and results. At first, we briefly discuss the experimental setup and the details of discriminative descriptor learning. Then, we compare our proposed color descriptor with several photometric color descriptors on three image data sets. Next,

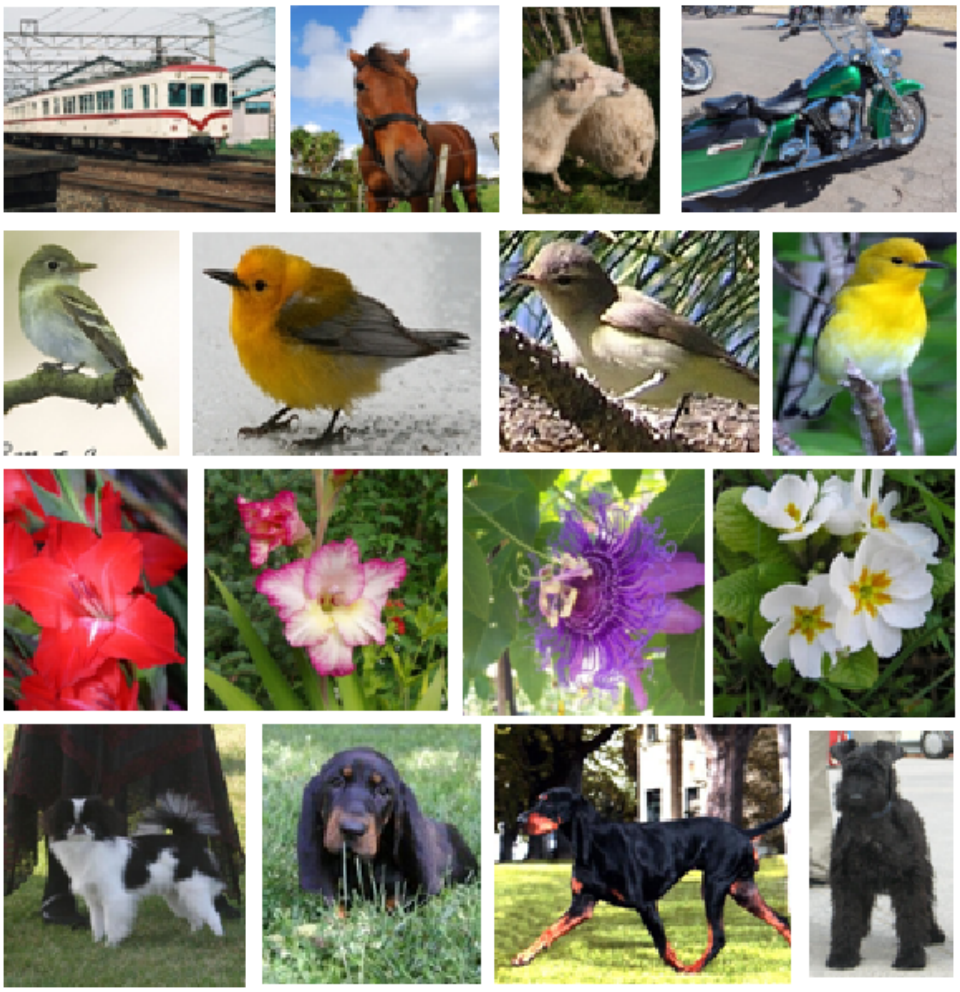

Figure 4.9: Example images from the four data sets used in this work. From top to bottom: PASCAL 2007, Birds-200, Flowers-102, Dogs-120.

we focus on the universality aspect of our descriptor and compare universality with specificity. In our final experiments, we combine our descriptor with shape description and compare results to the state of the art.

4.5.1 Experimental setup

In this section, we briefly discuss the experimental setup used for sections 4.5.2 and 4.5.3. For these two sections we use a comparatively simpler framework to reduce the computational time, as our goal is to assess relative performance. For both sections, we choose three challenging image data sets, namely, Flower102 [65], Birds200 [94] and PASCAL 2007 (Figure 4.9). For Flowers and Birds, the colors over the object classes are relatively con-

stant. However for PASCAL 2007, colors are likely to change significantly in between samples of the same class (consider e.g. cars). In these experiments, we use a regular dense grid (16×16) with 50% overlap to extract patches from the images. After description of the patches, we employ a K-means on a random subset of features from the training set to build the visual vocabulary. We use SVM with an intersection kernel to obtain the classification score. The training and test set selection is consistent with the corresponding cited articles for each data set. For section 4.5.4, we use a different experimental setup which is discussed in the beginning of that section.

For descriptor learning, for each data set we convert all the training images from sRGB to L*a*b and construct a 3D histogram quantizing the L*a*b space by $10 \times 20 \times 20$, then we convolve these 3D histograms using a gaussian filter ($sigma = 1$). They are then used as 4000 dimensional feature vectors. We adapt the DITC implementation from [19] and use the Graph Cut implementation from [23]. As discussed in section 4.3.2, there are two parameters in our descriptor learning, namely, the dilation and smoothness cost parameters. The dilation cost parameter should be ideally equal to infinity, so we use a large enough value for that. Empirically we found that a smoothing cost parameter $\alpha_D = 10^{-8}$ obtained satisfying results on all data sets, and kept it constant.

We compare the clusters computed with standard DITC to the clusters computed with our algorithm which enforces connectivity and smoothness of the clusters. In Figure 4.8, we can clearly see that our method produces connected and smooth clusters. Note that, non-connected green parts from the first two clusters are associated to the green cluster when our method is employed. DITC only concerns discriminative clustering and does not ensure connected clusters which is undesirable from a colorimetric point of view.

4.5.2 Discriminative color descriptors

The aim of this work is to arrive at a better color descriptors for object recognition directly on the discriminative power of the final representations. We start by comparing our discriminative descriptor(DD) to other pure color descriptors and the color name descriptor [87]. Note that in several comparisons color names were found to outperform various other pure color descriptors [36, 39].

We consider two well known photometric invariants: normalized RGB (rg histogram) and a hue histogram (HH) [3] and the Color Names(CN) [4] [87].

3. Implementation provided by K. van der Sande at http://koen.me/research/colordescriptors.

4. As a sanity check we performed a k-means based LAB descriptor. Results were found to be inferior.

Method	Flower102	Bird200	Pascal2007
rg	38.6%	4.3%	10.6%
HH	32.8%	3.5%	10.1%
CN	40.2%	7.7%	11.6%
DD(11)	43.7%	8.0%	12.2%
DD(25)	47.0%	8.7%	12.6%

Table 4.1: Comparison with photometric invariants.

We compare them against our descriptor with two settings, namely 11 and 25 clusters. Table 4.1 contains the experimental results. For each data set we show the classification accuracy (or mean average precision for PASCAL 2007). For the case of 11 dimensions (equal to the CN descriptor) our descriptor obtains improved results on Flower and Bird, but slightly lower results than color names on PASCAL 2007. We can see from the table that our descriptor with 25 dimensions outperforms all the other descriptors used in the experiment. Note, that it is unclear how to increase the dimensionality of the color name descriptor above the eleven basic color names.

4.5.3 Universality versus specificity

We discussed universality color descriptors because of their ease of use in section 4.4. In general, there is a growing interest in across-data set generalization of methods in the community [84]. Here we use again the three data sets. We follow a leave-one-out approach, where we learn our descriptor on two data sets and test on the other. We also do data set specific experiments, where we learn on one data set and test on the same. In each case, we learn 3 different cluster groups i.e. $k = [11, 25, 50]$ using our proposed method. We follow similar setup as section in 4.5.2 to represent images as bag-of-visual-words.

It is evident from Figure 4.10 that for larger k, the difference between universality and specificity becomes smaller. Also note that, the best results obtained using our universal descriptor, although not better than the specific ones, outperform other state-of-the art color descriptors used in experiments of section 4.5.2. In conclusion, for larger dimensions the drop of performance due to universality is relatively small, and users could prefer using it, rather than having to train a new data set specific descriptor.

4.5.4 Discriminative descriptors vs state-of-the-art

We compare our approach with the state-of-the-art approaches in the literature. The experiments are performed on Birds-200, Flowers-102 and PASCAL 2007. Additionally, we also show the applicability of our approach on the challenging Stanford-Dogs 120 data set. For our final experiments,

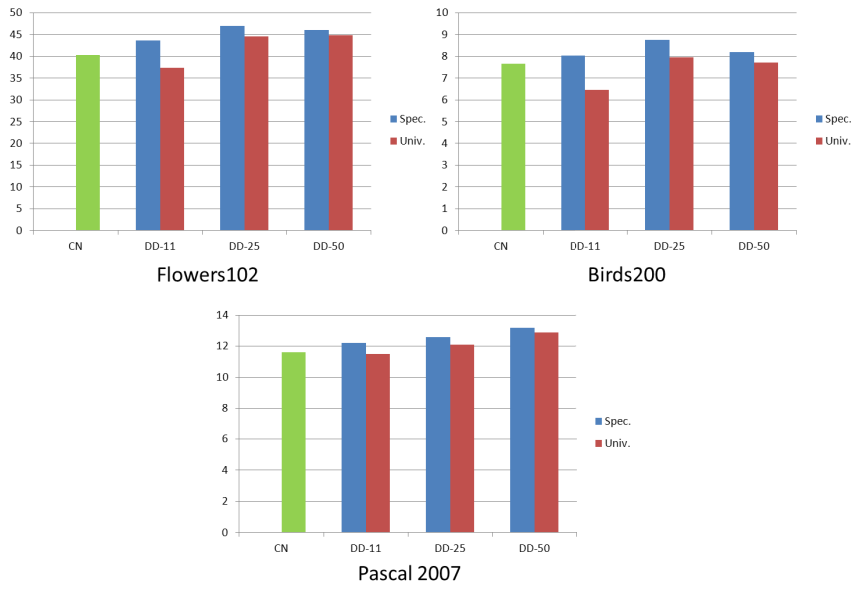

Figure 4.10: Universality versus Specificity. The green bar (the left bar of each plot) is the state-of-the-art pure color descriptor (Color Names).

we followed the standard bag-of-visual-words pipeline. For feature detection, we use a combination of multi-scale grid with interest point detectors. For shape we use the SIFT descriptor. A visual vocabulary of 4000 is constructed for shape representation. For color, we use a visual vocabulary of 500 words. The vocabularies are constructed using standard K-means and the histograms are constructed using hard assignment. To represent an image we use the spatial pyramid representation as in [50]. For classification, we use the non-linear SVM using the χ^2 kernel [104]. We also compare our approach with the ColorSIFT descriptors [86] on the PASCAL VOC 2007, Birds-200 and Flowers-102 data sets. We use CSIFT descriptor for the PASCAL VOC 2007 data set and OpponentSIFT for the other two data sets. A visual vocabulary of 4500 is constructed for ColorSIFT descriptors and an image is represented by spatial pyramids. The results are summarized in Table 4.2.

On the Birds-200 data set, shape alone provides a classification performance of 15.3. Our final result is a combination of late fusion between discriminative color and shape, shape alone and color alone. On this data set our discriminative approach achieves the best classification score of 26.7 outperforming the colorSIFT [86] based on the same detected features. The universal color names result in a slight drop in performance. The other approaches in Table 4.2 also use a combination of color and shape. The portmanteau approach employs both color and shape to learn a compact color-shape vocabulary. The tricos approach [11] uses segmentation tech-

nique whereas for image representation shape and color with fisher vectors are employed.

On the Flowers-102 data set, a mean accuracy of 69.0 is obtained. The incorporation of proposed color approach together with shape leads to 81.3. The universal color descriptor learned on the PASCAL 2007 and Birds-200 data set results in a slight drop in performance. On this data set again, our approach provides a comparable results to the state-of-the-art approaches in literature [10, 11, 37, 65]. On the PASCAL 2007 data set, our framework with shape alone provides a meanAP of 59.9. Adding color with shape increases the meanAP to 62.0. The universal color descriptor results in slight deterioration in performance with a meanAP of 61.7. Again on this data set, our final results are comparable to state-of-the-art results in literature [12, 39, 86, 107]. The method of [39] uses color attention approach to combine with color and shape with a meanAP of 58.0. The best reported results of 64.4 [107] is obtained using a different coding technique. Note that in this work we use the standard vector quantization with hard assignment. However, our color descriptor can be used in any encoding framework together with SIFT.

Finally, we have included the challenging Stanford Dogs 120 data set. This data set is interesting because dog furs only exist in a reduced set of colors (mainly browns, black and white). Here our approach provides a classification score of 28.1 compared to 21.1 using shape alone. On this data set, we use the shape features kindly provided by the authors. The universal color descriptor (learned from PASCAL, Birds and Flowers data set) results in a drop in performance to 26.5. From which we can see that for particular (in a color sense) data sets computing a specific color representation can still yield a large performance gain. To the best of our knowledge the final score of 28.1 obtained in this work is the best performance achieved on this data set in literature [10, 11, 43].

In summary, despite the simplicity of our approach we show that excellent performance can be achieved using a combination of color and shape. The applicability of proposed approach is apparent on wide range of data sets from Stanford-Dogs to PASCAL VOC 2007.

4.6 Conclusion

In this chapter, we have proposed a way to design discriminative color descriptors for image classification. By taking an information theoretic approach, our descriptor provides a certain degree of photometric invariance while maximizing the discriminative power. Interestingly, when the descriptor is learned on two data sets and test on a third one, the performances are close to those obtained when learned and tested on the same data set. This universal property is very attractive since users are not required to

Method	Birds-200	Flowers-102	Pascal 2007	Dogs-120
Tricos [11]	25.5	85.2	-	26.9
Bicos [10]	23.7	**85.5**	-	25.7
portmanteau [37]	22.4	73.3	-	-
Color Attention [39]	-	-	58.0	-
MKL [65]	-	72.8	-	-
LLC [43]	-	-	-	14.5
Fisher [12]	-	-	61.7	-
Super Vector [107]	-	-	**64.0**	-
Shape alone	15.3	69.0	59.9	21.7
ColorSIFT	20.4	77.6	57.4	-
This paper (universal)	26.3	79.4	61.7	26.5
This paper (specific)	**26.7**	81.3	62.0	**28.1**

Table 4.2: Comparison of state-of-the-art results with our approach. Note that our approach provides best results on two data sets. The results in the upper part of the table are obtained from the corresponding papers, the results in the bottom part of the table are obtained based on the same detected features.

learn a new color representation for every new data set and can just apply the universal color representation to their problem. Finally, we have shown that our color descriptors provide state-of-the-art results when combined with shape descriptors. Since the clustering step only exploits the global distribution of the colors in the images, future works will consist in accounting local spatial interactions between the colors during this step in order to adapt the clusters to local descriptors such as those used in the classical bag-of-visual-words.

Chapter 5

Towards Multispectral Data Acquisition with hand-held Devices

Résumé: Nous proposons une méthode d'acquisition de données multispectrales à l'aide d'un appareil portable possédant un écran et une caméra RGB. Le principe est d'utiliser l'écran comme source d'éclairage et la caméra pour capturer des images éclairées par les couleurs primaires rouge, vert et bleu de l'écran. La combinaison des trois éclairages et des trois fonctions de réponse de la caméra permet d'obtenir neuf réponses différentes qui sont utilisées pour l'estimation de la réflectance de la surface éclairée. Les résultats sont prometteurs et montrent que la précision de la reconstruction spectrale est améliorée d'un facteur 30% à 40% comparé à celle obtenue à partir d'un seul éclairage. D'autre part, nous proposons de calculer des fonctions de base adaptées à la combinaison capteur-éclairage en éliminant la partie de la réflectance qui tombe dans la zone aveugle de cette combinaison. Nous montrons expérimentalement qu'optimiser l'estimation de la réflectance sur ces nouvelles fonctions de base permet de réduire significativement la variance de l'erreur par rapport à des fonctions de base indépendantes de la combinaison capteur-éclairage. Nous concluons que l'acquisition de données multispectrales est potentiellement possible avec un appareil portable grand public comme une tablette, un téléphone ou un ordinateur portable, ouvrant ainsi la porte à des applications qui sont actuellement considérées comme irréalistes.

Abstract: We propose a method to acquire multispectral data with handheld devices with front-mounted RGB cameras. We propose to use the display of the device as an illuminant while the camera captures images illuminated by the red, green and blue primaries of the display. Three illuminants and three response functions of the camera lead to nine response values which are used for reflectance estimation. Results are promising and

show that the accuracy of the spectral reconstruction improves in the range of 30-40% over the spectral reconstruction based on a single illuminant. Furthermore, we propose to compute sensor-illuminant aware linear basis by discarding the part of the reflectances that falls in the sensor-illuminant null-space. We show experimentally that optimizing reflectance estimation on these new basis functions decreases the RMSE significantly over basis functions that are independent to sensor-illuminant. We conclude that, multispectral data acquisition is potentially possible with consumer hand-held devices such as tablets, mobiles, and laptops, opening up applications which are currently considered to be unrealistic. [1]

5.1 Introduction

The electromagnetic spectrum ranges from radio waves of wavelength $\lambda > 1$m to gamma rays with a $\lambda < 10^{-12}$m. The portion of this spectrum that can be directly observed by the human visual system(HVS) is incredibly small in comparison. The HVS is roughly responsive to light with wavelengths from 400nm to 700nm. It is this small spectral band, which defines how we see the world. Multispectral color is a function of these wavelengths. For many applications CIE tri-value description of color, as is provided by for example XYZ or $L^*a^*b^*$ values, is not sufficient and a multispectral description is desired. These applications vary from consumer products, such as paint selection, online cloth shopping, cosmetics industry and to more specialized fields such as in eHeritage and fruit quality assessment. For all these applications multispectral acquisition of color allows users to disentangle the set of metamers (different multispectral reflectances which map to the same tri-value), and provides a more precise description of color.

There exist two main approaches to multispectral data acquisition. The first method, and by far the most popular one is based on passing the light through filters which pass only part of the light. These filters can be a set of narrow band filters or sophisticated optical filters like AOTFs or LCTFs. Either by changing the narrow band filters over time or by splitting the light into different wavelengths, a multispectral measurement is acquired. A variety of such multispectral cameras exist in the market, and they have as main advantage that they are very accurate. However, because the acquisition of multispectral camera is expensive this equipment is only available to a few specialized laboratories.

A second approach to multispectral imaging is based on statistical learning [16, 29]. In this case, the multispectral data can be estimated from

1. The content of this chapter is accepted in the International Conference of Image Processing (ICIP), 2013 [41].

Figure 5.1: Multispectral data can be obtained with hand-held devices by using the screen to illuminate the object under various illuminations. The acquired measurements can be used to reconstruct the spectral reflectance.

relatively fewer number of measurements. Thus, the series of narrow band filters from the previous approach are replaced by fewer number of broad-band filters. A variation to this approach exists which is based on the duality between light sources and filters. However, the field of improving the spectral resolution by changing the illuminants (instead of the filters) has received relatively little attention [14, 70]. Park et al. [70] show that theoretically it is possible to obtain multispectral information from a camera by varying the illuminant. Furthermore, they experimentally show that in a highly controlled environment with high quality calibrated acquisition equipment it is possible to obtain multispectral information with a camera. The main advantage of these methods is that one does no longer require a multispectral camera (or chromatic filter set) to obtain multispectral images. The main drawback of this method is that it requires users to illuminate the scene with various illuminants. Because of the drawback, this approach to multispectral imaging has attracted relatively little attention. In a similar work, Chi et al. [14] propose to place the filters in front of the light source (instead of the lens) to generate multiple illuminations. One advantage of their approach is to be able to eliminate the effect of ambient light. However, requirement and selection of additional filters remains a problem in their approach as well.

In this chapter, we propose to use the screen of the hand-held devices to display a set of illuminants which can then be observed by a front-mounted camera. In Figure 5.1 an illustration of the proposed approach is given. The object with unknown reflectance is held in front of a device comprising a display and an integrated camera. At the same time that the camera captures images of the object, the display depicts different colors, and thereby changes the illumination of the object. This process results in a set of acquisitions of the reflectance under varying illuminants. The output of the system will be a multispectral signature of the object estimated from these measurements. The main originality of our approach lies in the functionality

shift where we use the display of the device as an illumination source. The observation that for many hand-held devices cameras are mounted alongside a display makes this an especially convenient solution. This is also true for laptops with a built-in web cam. Note that this solves the main drawback of earlier work [14, 70] which was the availability of multiple illuminants. Wandell and Farrell [90] proposed a method to use a scanner as a colorimeter. However, their method remains based on a tri-value input, thereby limiting the obtainable accuracy. For the spectral estimation, we propose an improvement on the method of Park et al. [70]. They use the observation that real-world reflectances can be well-approximated with a low-parameter linear model [71]. The basis functions for this low-parameter model are typically derived from the spectra of the Munsell color chips. In this work, we use basis functions that discard the part of the Munsell reflectances that falls in the sensor-illuminant null-space. We have shown that the new set of basis functions improves the accuracy of the spectral reconstruction. In the next section, we are going to describe the traditional way to multispectral imaging.

5.2 Multispectral color imaging

The most accurate way to obtain multispectral color is by using multispectral cameras. Multispectral cameras are sophisticated and expensive devices only available to specialized laboratories. The most simple multispectral cameras use a color wheel fitted with a set of narrow band color filters(Figure **??**). Color filters are used on image sensors to allow them to see color as humans do. These color filters are placed on or directly above the sensor to selectively filter the unwanted wavelengths and pass the desired ones. The filter wheel rotates and the camera captures one gray scale image per filter. Typically, 12 to 20 color filters are used but this number can vary across cameras. These kinds of cameras are used to take multispectral images of a stationary scene. Note that, to obtain reflectance of colored surface the scene illumination must be known. This is why, a white lambertian surface is almost always placed in the scene which gives information about the illumination.

There exist more sophisticated multispectral cameras that uses optical filters like ATOFs and LCTFs, but they work under the same principal of measuring multiple narrow bands of light. These type of cameras are able to provide very accurate measurements but their high cost is a major obstacle to overcome. This is why, over the years scientists have come up with techniques to extract multispectral information using fewer number of measurements with consumer RGB cameras thanks to statistical learning.

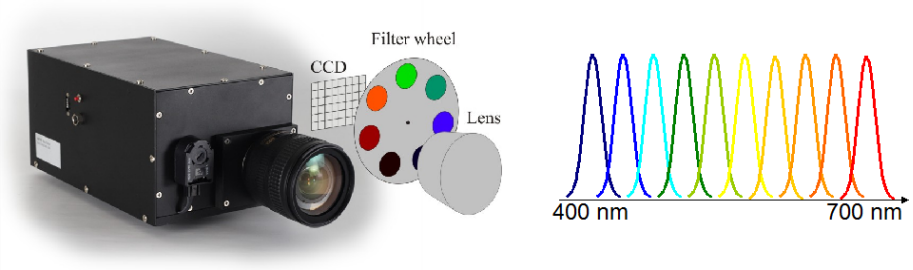

Figure 5.2: On the left: a multispectral camera with color wheels; on the right: narrow band filters typically used with this kind of multispectral cameras.

5.3 Multispectral reflectance estimation from RGB camera responses

For a typical multispectral camera, several measurements are required to obtain a multispectral image. However, for reflectance estimation from RGB cameras only a few measurements are taken. For each measurement, a RGB camera provides 3 responses (i.e. R,G,B), so for T measurements we have $3T$ camera responses. Generally, filters are used to obtain multiple measurements. However, as explained before, one could also take these measurements varying the illumination. The mathematical formulation of the problem in both cases (varying filters or varying illumination) is identical. Image formation in a camera can be modeled as

$$\rho_{mn} = \Gamma\left(\int_{\lambda} r(\lambda)c_m(\lambda)s_n(\lambda)\mathrm{d}\lambda\right) + \epsilon_m \tag{5.1}$$

where, ρ_{mn} is the response of the camera. Γ is the camera non-linearity, $r(\lambda)$ is the reflectance of a point in the object, $c_m(\lambda) : m = 1 \ldots M$ is the spectral sensitivity of the m-th channel of the camera, M being the total number of sensors and $s_n(\lambda) : n = 1 \ldots N$ is the spectral power distribution (SPD) of the scene illumination, N being the total number of illuminations. We denote the total number of measurements as $\mathcal{N}$, where, $\mathcal{N} = MN$. Finally, ϵ_m is the noise of the m-th channel of the camera.

The main challenge is to reconstruct the reflectance $r(\lambda)$ of the object from these $\mathcal{N}$ measurements obtained from the camera. If we consider sampling the spectra into $\mathcal{I}$ discretized λ, by merging the illuminations and the camera sensors into one matrix $\boldsymbol{F}$ where each column is the product between one sensor sensitivity and one illuminant SPD, Eq. 5.1 can be rewritten in matrix form as:

$$\boldsymbol{\rho} = \left(\boldsymbol{F}^T\boldsymbol{r}\right) + \boldsymbol{\epsilon} \tag{5.2}$$

where $\boldsymbol{\rho}$ is the response of all the sensor-illumination pairs given the reflectance $\boldsymbol{r}$. Both $\boldsymbol{\rho}$ and $\boldsymbol{r}$ are column vectors. If RAW images from the

camera are available, the camera non-linearity factor could be removed. The goal is to estimate $\boldsymbol{r}$ from the Equation 5.2 when $\boldsymbol{\rho}$ is known ($\boldsymbol{F}$ could be known or unknown). The problem is known as reflectance estimation problem. Estimating $\boldsymbol{r}$ from Equation 5.2 leads to an ill posed problem. So, very often additional constraints are needed to solve the problem. Note that, in our notations, we have used bold faces for vectors and matrices.

5.3.1 Reflectance estimation

In the following sections, we shortly discuss some popular reflectance estimation methods which we will compare in the experimental section. These methods introduce different constraints to compute the reflectance $\boldsymbol{r}$ from Equation 5.2. Among a series of successful reflectance estimation algorithms proposed during the last two decades, one category of methods works under the principle of learning a mapping between the camera measurement space and the reflectance space [16, 32]. This type of methods does not require any statistical *a priori* information but involves a laborious training phase. Another category of methods, estimates spectra using statistical *a priori* information [29, 70]. This category of methods makes assumptions about the spectral space and often require the knowledge of the illuminant-sensor matrix $\boldsymbol{F}$, however, training phase for these methods are simpler. Methods from both the categories have shown excellent results in the past. Thus, in the next section, we present two of the popular methods from each category.

5.3.1.1 Reflectance estimation without statistical *a priori* information

This group of methods estimates the reflectance spectra from the measurements using a collection of reflectance/measurement pairs. Assuming we have a labeled training set, $S_{tr} = ((\boldsymbol{\rho}_1, \boldsymbol{r_1}), \ldots, (\boldsymbol{\rho}_{\mathcal{L}}, \boldsymbol{r}_{\mathcal{L}}))$ of measurement vectors $\boldsymbol{\rho_j}$ and corresponding reflectance spectra $\boldsymbol{r_j}$. In matrix form, we denote the measurement matrix as $\boldsymbol{\rho}_{tr} = [\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \ldots, \boldsymbol{\rho}_{\mathcal{L}}]^T \in \mathbb{R}^{\mathcal{L}\times\mathcal{N}}$ and reflectance matrix as $\boldsymbol{R}_{tr} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{\mathcal{L}}]^T \in \mathbb{R}^{\mathcal{L}\times\mathcal{I}}$, $\mathcal{N} < \mathcal{I}$. In most of the cases, $\boldsymbol{R}_{tr}$ is a set of Munsell spectra. This category of methods can account for the camera noise directly. Moreover, they do not require the knowledge of the matrix $\boldsymbol{F}$ in Equation 5.2 as they directly learn a mapping from camera responses $\boldsymbol{\rho_{tr}}$ to the reflectance $\boldsymbol{R}_{tr}$. On the down side, these methods require the knowledge of camera responses for the training reflectances to obtain S_{tr}, which is often a laborious task.

Direct pseudo-inverse solution: First used by Day et al [16], direct pseudo-inverse is the simplest reflectance estimation technique existing till date. It learns a direct linear mapping from camera measurement space to reflectance space from the camera responses of some known spectra (e.g.

Munsell Spectra). Then, given new camera measurements $\boldsymbol{\rho}_{ts}$ from an unknown material, we can estimate its reflectance $\boldsymbol{r_{ts}}$ as:

$$\boldsymbol{r}_{ts} = \boldsymbol{R}_{tr}^{T}\boldsymbol{\rho}_{tr}(\boldsymbol{\rho}_{tr}^{T}\boldsymbol{\rho}_{tr} + \gamma\boldsymbol{I}_{\mathcal{N}})^{-1}\boldsymbol{\rho}_{ts} \tag{5.3}$$

where γ is the regularization parameter and $\boldsymbol{I}_{\mathcal{N}}$ is an identity matrix of $\mathcal{N} \times \mathcal{N}$ dimension.

Kernel Regression: Kernel regression projects the measurements to a higher dimensional Hilbert space and learns a mapping from this new space to the spectra space. This allows to realise non-linearities in the mapping. Equation 5.4 presents the kernel regression for reflectance estimation [32]

$$\boldsymbol{r}_{ts} = \boldsymbol{R}_{tr}^{T}(\boldsymbol{K}_{tr} + \gamma\boldsymbol{I}_{\mathcal{L}})^{-1}\boldsymbol{\kappa}^{\boldsymbol{\rho}_{ts}} \tag{5.4}$$

Here, ($\boldsymbol{K}_{tr} = \boldsymbol{\rho}_{tr}^{\Phi}{\boldsymbol{\rho}_{tr}^{\Phi}}^{T}$ and $\boldsymbol{\kappa}_{\boldsymbol{\rho}^{ts}} = \boldsymbol{\rho}_{tr}^{\Phi}\boldsymbol{\rho}_{ts}^{\Phi}$), where $\boldsymbol{\rho}_{tr}^{\Phi} = [\Phi(\boldsymbol{\rho}_1), \Phi(\boldsymbol{\rho}_2) \ldots \Phi(\boldsymbol{\rho}_{\mathcal{L}})]^{T}$, $\boldsymbol{\rho}_{ts}^{\Phi} = \Phi(\boldsymbol{\rho}_{ts})$ and Φ is a function that projects the measurements to a higher dimensional space. In practice, $\boldsymbol{K}_{tr}$ and $\boldsymbol{\kappa}_{\boldsymbol{\rho}^{ts}}$ are calculated using a distance metric and with the true measurements $\boldsymbol{\rho}_{tr}$ and $\boldsymbol{\rho}_{ts}$. $\boldsymbol{I}_{\mathcal{L}}$ is an identity matrix of $\mathcal{L} \times \mathcal{L}$ dimensions. In Heikkinen et al. [32] the authors show that even for essentially linear systems non-linear mapping can improve estimation accuracy over a linear one.

5.3.1.2 Reflectance estimation using *a priori* information

This category of methods manipulates *a priori* information to improve the estimation accuracy. Lawrence Maloney [60] have shown that higher dimensional spectral reflectance could be represented in a lower dimensional space without important loss of information. This lower dimensional space is usually obtained by applying PCA on a large set of natural spectra (e.g. Munsell spectra). Their work suggests that, all natural spectra only occupy a sub-space of the higher dimensional spectral space. The spectral estimation method described in this section use this information as statistical *a priori* to solve the estimation problem. Moreover, this family of reflectance estimation algorithms require the knowledge of the camera sensitivity curve, the illumination spectral power distribution and/or filter responses. However, they do not require a training set to train the system. We present two reflectance estimation techniques which fall under this category.

Wiener estimation [29]: This method solves the estimation problem under the assumption of normally distributed data. The covariance matrix calculated from the Munsell spectra and the sensor-illuminant pair are used to compute the mapping from camera responses to the reflectance. The formulation of Wiener estimation is as follows:

$$\boldsymbol{r_{ts}} = \boldsymbol{\Sigma}_{RR}\boldsymbol{F}(\boldsymbol{F}^{T}\boldsymbol{\Sigma}_{RR}\boldsymbol{F} + \gamma\boldsymbol{I}_{\mathcal{N}})^{-1}\boldsymbol{\rho_{ts}} \tag{5.5}$$

where, Σ_{RR} is the covariance matrix calculated from the Munsell spectra.

Park et al [70]: This method has been discussed briefly in the previous section. We propose to improve the method of [70] due to the similarity of their work with ours (multispectral reconstruction from RGB image taken under varying illuminants). In the next section we detail this method and propose our modification to improve it.

5.4 Multispectral imaging by varying illumination

In this section, at first we present the spectral estimation method proposed by [70]. Then, we proposed our modification to this method to improve the accuracy.

5.4.1 Reflectance estimation by optimization of low-parameter representation [70]

In [70], the author use a led-based customized illumination system to facilitate fast capture of multispectral data. To reconstruct the reflectance, they propose to minimize a regularized constrained optimization problem. They project the sensor-illumination pairs in a low parameter orthogonal basis function space and optimize basis vector coefficients to minimize the $L2$-norm in the measurement space.

Park et al. [70] approximate the reflectance of real-world materials with a limited number of spectral basis functions according to:

$$r(\lambda) \approx \sum_{k=1}^{K} \sigma_k b_k(\lambda) \tag{5.6}$$

where σ_k are scalar coefficients of the k-th basis function $b_k(\lambda)$. These basis functions can be computed with eigenvector analysis (PCA) of the 1257 Munsell color chips [71]. First four basis vectors computed from Munsell color chips are shown in figure 5.4.

Equation 5.6 could be writtent in Matrix form as:

$$\boldsymbol{r} = \boldsymbol{B}\boldsymbol{\sigma} \tag{5.7}$$

where $\boldsymbol{\sigma}$ is a column vector of the coefficients and $\boldsymbol{B}$ is a matrix whose columns are the basis functions. Substituting $\boldsymbol{r}$ from Equation 5.7 into Equation 5.2 and considering we have RAW output from the camera with limited noise, we can rewrite Equation 5.2 in matrix form as following:

$$\boldsymbol{P}\boldsymbol{\sigma} = \boldsymbol{\rho}; \tag{5.8}$$

where $\boldsymbol{P} = \boldsymbol{F}^T\boldsymbol{B}$ i.e. we assimilate the basis functions $\boldsymbol{B}$ and sensor-illumiannt sensitivity $\boldsymbol{F}$ together in a single matrix $\boldsymbol{P}$. When $\boldsymbol{P}^T\boldsymbol{P}$ is invertible, we can get a least squares solutions $\boldsymbol{\sigma} = (\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T\boldsymbol{\rho}$. However,

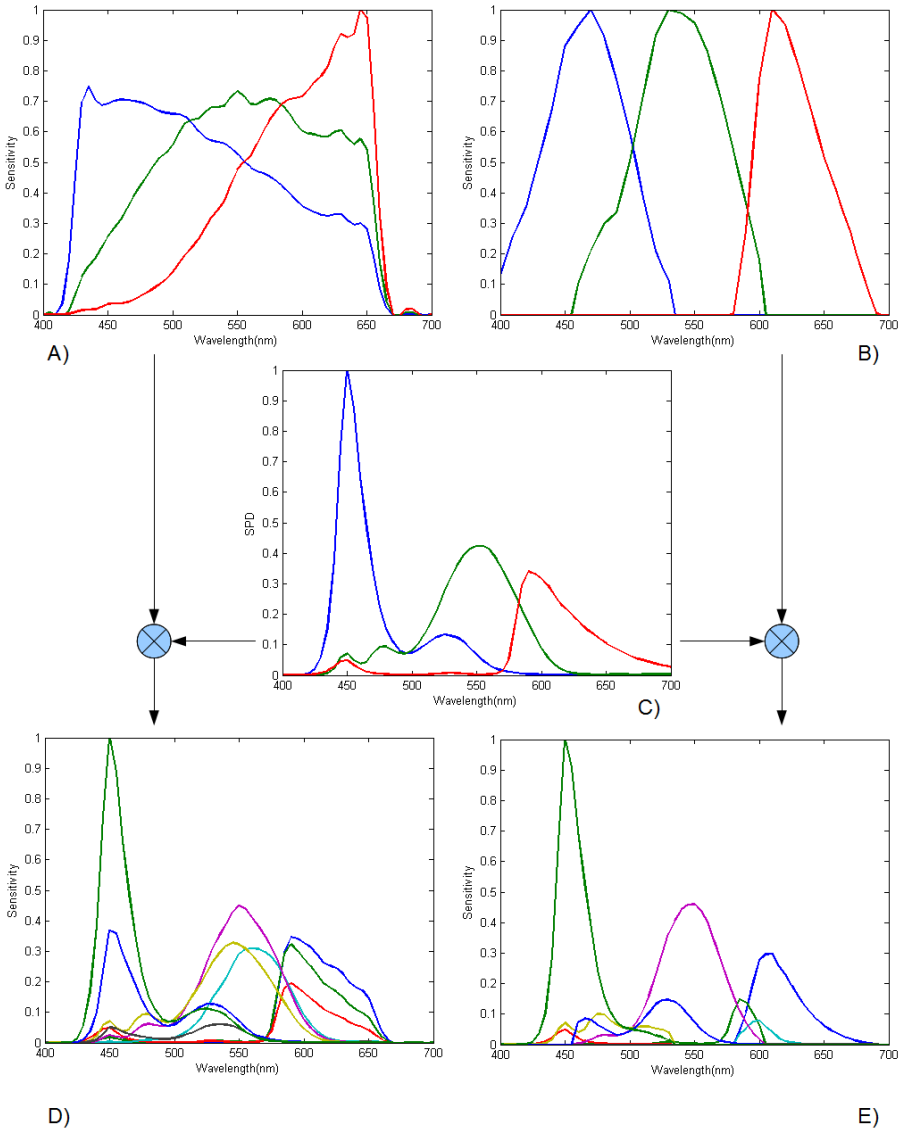

Figure 5.3: Sensor-illuminant pairs (D and E) resulted from two camera sensors of A) Sigma SD-10 and B) Retiga camera and C) RGB primaries of a DELL E4310 laptop. All the responses are normalized.

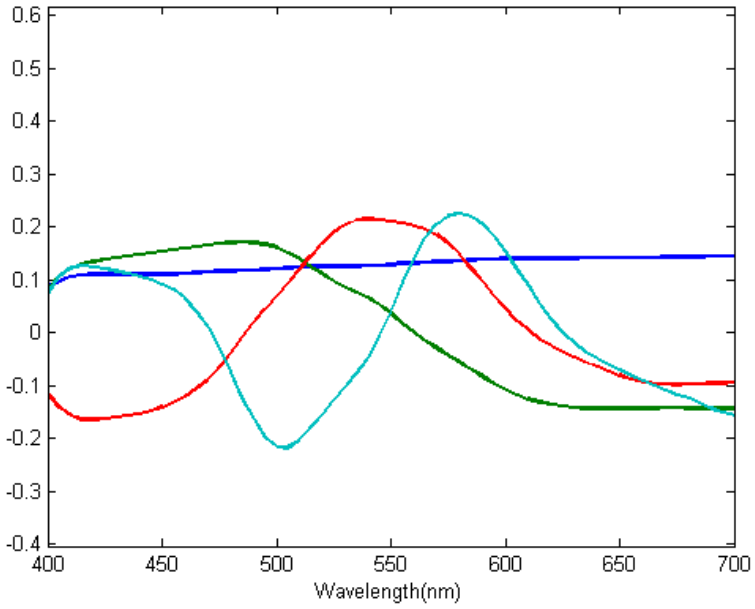

Figure 5.4: The first 4 basis functions calculated from the Munsell color chips using PCA. In this work, we use the first 8 eigenvvectors.

this least square solution may return some negative values. To prevent that, a positivity constraint is needed. At this point, the problem in hand can be formulated as a constrained minimization problem as follows:

$$\arg\min_{\sigma^+} |\boldsymbol{P\sigma} - \boldsymbol{\rho}|, \text{ subject to } \boldsymbol{B\sigma} \geq 0. \tag{5.9}$$

In Equation 5.9, the positivity constraint $\boldsymbol{B\sigma} \geq 0$ ensures that the reconstructed reflectance would not have negative values. The solution to the constrained quadratic minimization problem in Equation 5.9 may be numerically unstable if $\boldsymbol{P}$ has a large condition number. Thus, an additional constraint is required for a reasonable solution. To this end, Park et al. [70] propose to add a smoothness constraint since natural spectra tend to be smooth. This could be done by penalizing large values for the second derivative of the spectral reflectance with respect to λ:

$$\arg\min_{\sigma^+_{smooth}} \left(\left| \boldsymbol{P\sigma} - \boldsymbol{\rho} \right|^2 + \alpha \left| \frac{\delta^2 r\left(\lambda\right)}{\delta\lambda^2} \right|^2 \right) \text{ subject to: } \boldsymbol{B\sigma} \geq 0 \tag{5.10}$$

where α is the smoothness parameter and $\frac{\delta^2 r(\lambda)}{\delta\lambda^2}$ is the smoothness constraint. This constraint helps to obtain reasonable solution if the matrix $\boldsymbol{P}$ is ill-conditioned. Park et al. [70] propose to solve the optimization problem presented in Equation 5.10 for spectral reconstruction from camera

responses. In the next section, we present our proposal to modify Equation 5.10 with a set of sensor-illuminant aware basis functions to improve the reconstruction accuracy.

5.4.2 Reconstruction using sensor-illuminant aware basis functions

Differing from the existing methods, where the illuminants are chosen to optimize spectral resolution, we propose to use the screen of hand-held devices as the changing illuminant of the scene. As the illuminants we use the three primaries of the screen in isolation, giving a red, green and a blue illuminant. In the case of a RGB camera we therefore have a total of nine measurements: the three camera channels for each of the three illuminants. Sensor-illumination pairs obtained from two different cameras and a screen primaries are displayed in Fig. 5.3.
The method proposed by Park et al. [70], constraints the reflectance estimation problem by noting that the spectra of real-world reflectance can be well approximated by a low-parameter linear model [71]. This linear model is independent of the sensor system. Here, we investigate adapting the linear model to the sensors to achieve improved reflectance reconstruction. Since the approach proposed in [70] is independent of the sensor-illuminant sensitivities $\boldsymbol{F}$, there is no reason that the resulted basis functions are the best to reconstruct multispectral data from the given sensor-illuminant pairs. Moreover, in our case display illuminants and sensor sensitivities are intrinsic properties of the hand-held device. Thus, we propose to adapt the spectral basis to the sensor-illuminant spectra. Considering the reflectance of the Munsell data set $\boldsymbol{R}$, we propose to break the Munsell data set into two parts:

$$\boldsymbol{R} = \boldsymbol{R}^{\boldsymbol{F}} + \boldsymbol{R}^{\perp \boldsymbol{F}}, \tag{5.11}$$

where $\boldsymbol{R}^{\boldsymbol{F}}$ is the part of the spectra which is in the illuminant-sensor space and $\boldsymbol{R}^{\perp \boldsymbol{F}}$ is the part of the Munsell spectra which is perpendicular or in the null-space of the sensor-illuminant space [3] and for that reason cannot be observed. The matrix $\boldsymbol{R}^{\boldsymbol{F}}$ can be computed with:

$$\boldsymbol{R}^{\boldsymbol{F}} = \boldsymbol{R}\boldsymbol{F}(\boldsymbol{R}\boldsymbol{F})^{+}\boldsymbol{R}, \tag{5.12}$$

where $(\boldsymbol{R}\boldsymbol{F})^{+}$ is the Moore-Penrose pseudo inverse. Equation 5.12 projects the spectra $\boldsymbol{R}$ into the sensor-illuminant space to discard the part $\boldsymbol{R}^{\perp \boldsymbol{F}}$ and subsequently back-project the projected values into the spectral space using a direct-pseudoinverse method. The obtained set of spectra $\boldsymbol{R}^{\boldsymbol{F}}$ is the part which is not orthogonal to the sensor-illuminant space. Then we propose to apply a PCA on the $\boldsymbol{R}^{\boldsymbol{F}}$ reflectances rather than on the original $\boldsymbol{R}$ reflectances. In this way, the resulted basis functions are not disturbed by

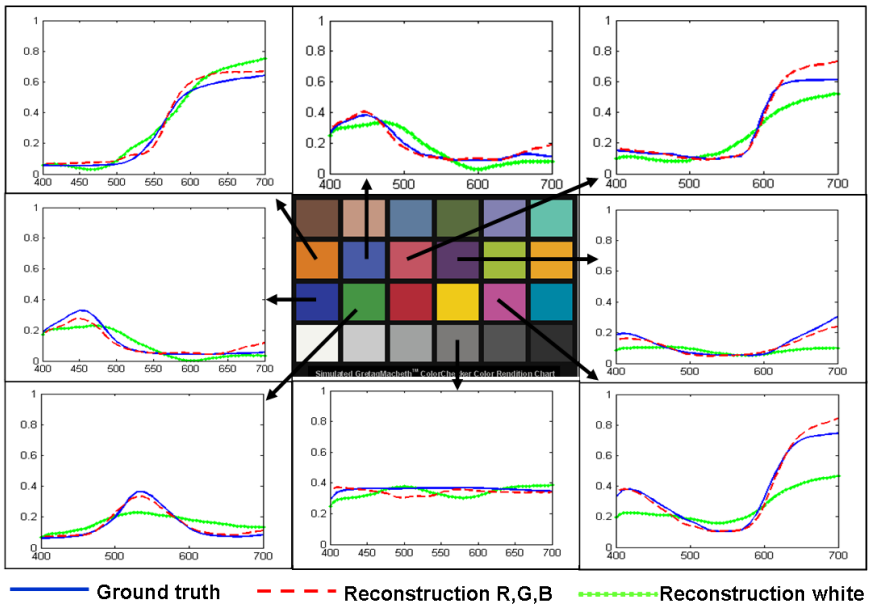

Figure 5.5: Comparative spectral estimation between R,G,B and white illuminants.

information that is not visible from the acquisition device given the sensor-illumination sensitivities. We denote these sensor-illuminant aware basis functions as $\boldsymbol{B'}$. Consequently, we denote the matrix $\boldsymbol{P}$ in Equation 5.8 as $\boldsymbol{P'}$ when computed from this new basis functions. Finally, in order to estimate the spectral reflectances from the camera responses $\boldsymbol{\rho}$, we have the following minimization problem:

$$\underset{\sigma^{+}_{smooth}}{\arg\min} \left(\left| \boldsymbol{P'\sigma} - \boldsymbol{\rho} \right|^2 + \alpha \left| \frac{\delta^2 r\left(\lambda\right)}{\delta\lambda^2} \right|^2 \right) \text{ subject to: } \boldsymbol{B'\sigma} \geq 0 \tag{5.13}$$

Practically, we use the Matlab function *quadprog* to solve Equation 5.13.

5.5 Experiments

In this section, we present experimental results on both synthetic data and real camera output. In each case, we compare the improvement of the estimation accuracy when using three i.e. R, G and B display illuminants over white i.e. R+G+B (the sum of the three display primaries). We also compare our method with [70]. We use two different metrics for the comparisons, namely, RMSE and CIEDE00 color difference [58].

Experimental results for Display Illuminant and Sigma SD-10 Camera								
Method	ColorChecker 24				ColorChecker 240			
	White		R,G,B		White		R,G,B	
	RMSE	CIEDE00	RMSE	CIEDE00	RMSE	CIEDE00	RMSE	CIEDE00
Pseudo-inverse	0.041	1.1	**0.014**	0.25	0.024	0.79	**0.0087**	0.23
Kernel Regr.	**0.037**	1.26	0.02	0.46	**0.020**	**0.77**	0.012	0.46
Wiener	0.0431	0.97	0.015	**0.24**	0.024	0.81	0.008	**0.22**
Park [70]	0.044	2.22	0.023	0.43	0.032	2.99	0.016	0.42
Our Method	0.042	**0.91**	**0.014**	0.29	0.025	0.98	0.010	0.34
Experimental results for Display Illuminant and Retiga Camera								
Pseudo-inverse	**0.039**	2.06	**0.01**	0.27	0.023	1.69	**0.008**	0.27
Kernel Regr.	0.0436	2.23	0.02	0.46	**0.020**	**0.77**	0.012	0.46
Wiener	0.04	**1.97**	0.012	**0.26**	0.023	1.67	0.0074	**0.25**
Park [70]	0.043	3.71	0.023	0.57	0.031	3.25	0.017	0.52
Our Method	**0.039**	2.26	**0.01**	0.31	0.024	1.89	**0.008**	0.32

Table 5.1: Comparison of reflectance estimation accuracy between R,G,B and white illuminants.

5.5.1 Experimental setup

In our experiments, a DELL Latitude E4310 laptop (13.3 inch screen size) is used as illuminant. LCD display technology is most commonly used in all the hand-held devices and laptops likewise. For capturing the image, we use the Sigma SD-10 camera which allows storing images in RAW format. We use a separate camera and not a camera mounted on a hand-held device because RAW shooting is still rare in hand-held devices available now, but considering the boom in the tablet computers and the mobile phone industry, we believe this to be common within a few years time. Like several other methods e.g. [29, 70], our proposed modification, requires the knowledge of sensor-illuminant sensitivities which could be obtained from the manufacturer in an ideal situation. In our case, we measure the laptop SPDs using a Konica Minolta CS-1000a spectroradiometer. The camera response curve for the sigma SD-10 camera was obtained from [1]. Moreover, for synthetic experiments, we use an additional camera sensor response curve of a retiga scientific camera to show the robustness of our method for different sensors. For kernel regression, we use a gaussian kernel following [32].

5.5.2 Synthetic data

In this section, we use the laptop display illuminant and two camera response curves. The Munsell color book spectra obtained from [2] are used for training and two Gretag Macbeth ColorCheckers of 24 and 240 colors are used for testing.

Table 5.1 shows the theoretical limit of the estimation performance for the given sensor-illuminant pair. We can see from the table that, for each algorithm, color checker and sensor-illuminant pair, the overall accuracy improves significantly when R,G,B primaries of the screen is used over the white (R+G+B). This significant improvement validates that multispectral

acquisition using hand-held device is a worthwhile proposition.
Comparison among different algorithms shows that, in most of the cases the regression based methods outperform the other methods. But in a real world scenario, the regression based methods involve a laborious process of image capturing as to train the system one must capture the image of all the train spectra. This is why, regression based methods lacks usability. On the other hand, [29, 70] and our proposed method require the knowledge of the sensor-illuminant sensitivities but does not require any other prior tasks to train the system thus make these methods more usable in the real-world context. For this reason, we only use these 3 methods in the next section for the experiments with real-world camera data. Note that, in most cases our proposed method significantly improve over the baseline method of Park et al. [70]. In some cases, our method provide best results among all the other methods including the regression based methods.

5.5.3 Real camera output

Here, we experimentally verify the accuracy of multispectral measurements which are obtained by illuminating materials with the primaries of a hand-held device.
Image acquisition: In this section, we use only the 24 patch color checker for reconstruction. LCD display technology is most commonly used in all the hand-held devices. The intensity (and to some extent the chromaticity) of an LCD display is sensitive to the angle of viewing. Moreover, because of this selective directional change of intensity, the illumination non-uniformity can be prevalent if the display is too close to the object of interest. Also the proximity of the object and the display screen eventually means the proximity of the object and the camera and may end up in sensor saturation. So, care must be taken while capturing images using a hand-held device for satisfactory results. As a rule of thumb, we hold the device orthogonal with the line that connects the plane of the display screen and the object plane. The distance between the device and the object is set to be approximately three-times the screen size. We assume that the camera and the scene are static. The effect of ambient light in reconstruction was not taken into account for this work, so the images are captured in a dark room with minimal ambient light. Ambient light could be taken into account as is shown in [14]. The comparison procedure is identical to that of the previous section. So, we capture 4 images of the color checker illuminated by R,G,B and the white(R+G+B).
Results: Table 5.2 shows the performance gain if three display primaries are used as illuminants over just one single light. For all the methods, the gain over white light is significant. The average RMSE gain on 24 patches varies from 30% for Park [70] to 40% for our method. CIEDE00 color difference is also improved significantly in each case. Moreover, our method outperforms

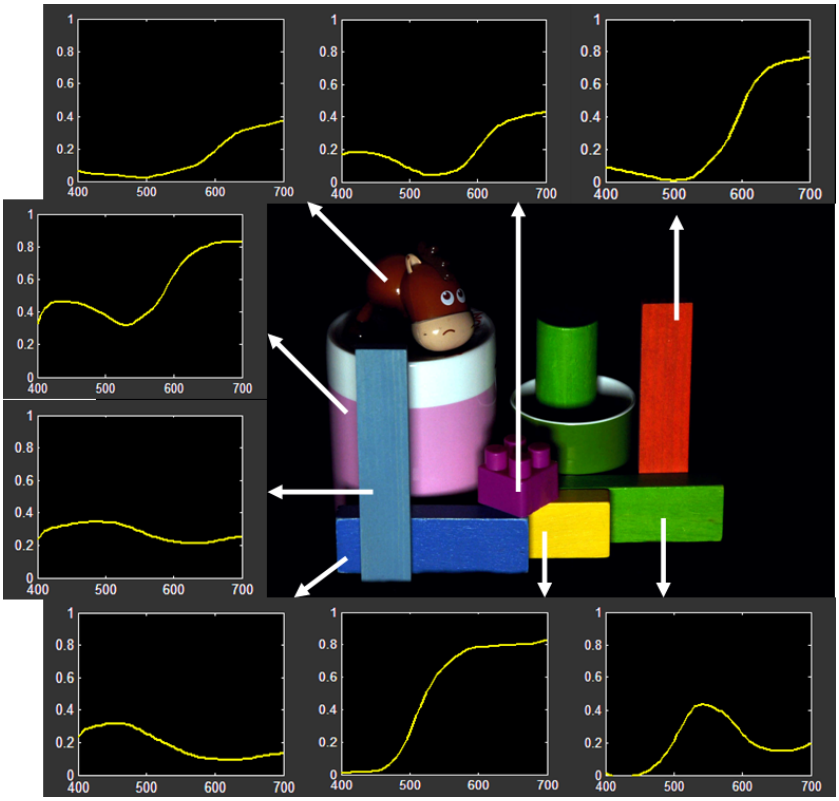

Figure 5.6: Spectral reflectance obtained by our method for several individual objects.

both Park et al. [70] and Wiener when R,G,B illuminants are used.
Figure 5.5 shows estimation results for some spectra of the color checker. It is evident that the use of various illumination helps to improve the spectral resolution and better estimate the sharp changes in the reflectances than the white light which provides a poor spectral resolution. As a further illustration of our method can be used to estimate the spectral reflectances of every pixel of a scene, which then can be used for relighting or color constancy. Figure 5.6 shows an example with several estimated reflectances spectra.

5.6 Application

Our proposed method allows the owner of a hand-held device to obtain accurate color information. It may not be obvious but accurate color measurement has many applications in everyday life. In this section, we give some example application to provide some insight in the tremendous potential of the proposed idea.

Method	White		R,G,B	
	RMSE	CIEDE00	RMSE	CIEDE00
Wiener	0.066	7.51	0.039	3.15
Park [70]	0.063	**7.4**	0.043	2.62
Our Method	**0.062**	**7.4**	**0.037**	**2.47**

Table 5.2: Reflectance estimation performance comparison between R,G,B and white illuminants for [70] and the proposed method

Application example 1: Consider an online customer interested in buying shoes in the same color as her dress. She holds the dress in front of the hand-held device and the multispectral analysis is performed. With this information the Internet retailer can now present the customer with matching shoes, as well as with other shoes which are considered to aesthetically match the color. Note that buying cloths based just on the color sensation of a single image is dependent on many scene accidental factors and can easily lead to unsatisfactory purchases.

Application example 2: A second application is in the paint industry. The customer is consider painting a wall in her kitchen in the same color as the wall in her living room. Finding the correct paint is known to be an arduous problem. After making a multispectral acquisition with her hand-held device and uploading the spectrum to the paint-companies web page, she could receive a much more precise advice on what paints to buy or how to mix them.

Application example 3: Another interesting application could be in the diagnosis of skin diseases. Color is an important clue to understand the type and severity of many skin diseases including skin cancer. To this end, multispectral color based diagnosis apps can help a patient to self diagnosis his/her disease.

5.7 Conclusion

We proposed a method which allows owners of hand-held devices with front-mounted RGB cameras to acquire multispectral data. Experiments show that potentially multispectral data acquisition with hand-held devices can significantly improve compared to taking a single color measurement under a known white light. In addition, our proposed algorithm improve reconstruction results on CIEDE00 scores up to 60%. However, current experiments did not include ambient illumination. Considering measurements in the presence of ambient light is one of the future direction. This work opens up multispectral data acquisition to all owners of hand-held devices with front-mounted cameras. The divulgation of this skill from the expert to

virtually everybody can open new applications in a number of fields such as the online retailing market, eHeritage, material recognition and biometrics.

Chapter 6

Conclusion and future works

6.1 Contributions

In this thesis, we have worked on spatial and color information aware image representation. In the chapter 3 and 4, we have proposed new image representations techniques and analyzed their performances on category-level classification task using the BoVW method. On the other hand, in the chapter 5, we proposed an innovative idea to acquire multispectral reflectance information of a surface using hand-held devices (smartphones, laptops, ipads) which could be useful for many computer vision applications.

In the chapter 3, our first contribution was to show that pairwise spatial relationships between visual words is an important information for category level recognition. Additionally, we have shown that, this information is complementary to SPR and thus could be combined with this method to improve the overall accuracy. We have proposed an original spatial encoding technique denoted Soft Pairwise Similarity angle distance histogram (SPS_{ad}) based on the concept of soft similarity between descriptors. We have shown that, the distribution of similar interest regions of the images is discriminative and can improve the performance of BoVW method significantly. We have also shown that, soft similarity is more robust and powerful than its 'hard' counterpart. The SSP_{ad} approach improved classification accuracy up to 16% on caltech101 and 7% on 15Scene data sets over the BoVW representation. When combined with SPR, the combination descriptor has shown to perform better than SPR and a group of other similar representations combined with SPR. In this case, our method improved the accuracy by 3.5% on Caltech101 and 2.5% on 15Scene data sets over the SPR method. In the chapter 4, we have proposed a novel approach to color description. We have proposed to learn the color descriptor from the training set rather than hand designing it. Our proposed descriptor is designed to be discriminative, contrary to most of the existing color descriptors that are designed

to be invariant. In our work, we presented specific and generic color descriptor. The specific descriptor is learned from the training images of a data set and tested on test images of the same data set whereas the generic descriptor work on a cross-data set principle. We have shown that, both specific and generic version of the descriptor outperform the state-of-the-art color descriptors on a classification task by as high as 7%. We also compared the performance of the color descriptor combining with the SIFT descriptor and compared it with the state-of-the-art results. In this case, our descriptor outperformed the state-of-the-art results on two out of four data sets and provided better results than the colorSIFT descriptor in all cases.
In the chapter 5, we looked at the problem of multispectral reflectance acquisition with hand-held devices, the innovative idea that features a functionality shift of the display of a hand-held device. In this work, the first contribution was the elimination of sophisticated setup for multispectral imaging by replacing it with hand-held devices. The second contribution was the improvement of an existing spectral estimation algorithm using a set of sensor sensitive basis functions. We have shown that, multispectral acquisition using hand-held devices is potentially possible and multispectral estimation from multiple measurements taken under R,G,B display primaries of hand-held devices can significantly improve compared to taking a single color measurement under a known white light. We have presented experiments on both synthetic data and real camera output. In all cases, multiple measurement improved accuracy over a single measurement. Moreover, our proposed method performs the best in case of real camera output. For multiple measurements the RMSE was improved by 65% over a single measurement. Moreover, our method improved RMSE by 16% over the method of [70].

6.2 Future Works

This thesis opens up a number of possibilities as future directions. In chapter 3, spatial relationships of pairwise visual words for improved BoVW representation has shown excellent performances. In this work, we use a basic (k-means and hard assignment) BoVW representation. One of the future directions could be to extend this work to a more recent BoVW coding techniques like sparse coding [98] or fisher vectors encoding [73]. Additionally, cross cue (color, shape) spatial relationship is also a promising direction.
The color descriptor learning approach presented in chapter 4 currently does not take into account the shape descriptors in use, even though color is almost always used in conjunction with shape to obtain the state-of-the-art accuracy. To this end, the discriminative descriptors could be learned with an additional constraints of shape vocabularies in use. Also, a soft discriminative color description is an interesting idea for future works.

The work on multispectral acquisition with hand-held device is very promising. However, the existing approach does not take into account the ambient light which is essential to increase the usability of the approach. Although, Chi et al. [14] has proposed a way to take the ambient light into account, their approach is not directly extendable in our settings due to the limitations of number of illuminants and the low luminance of the display. However, the most important follow up work should indeed test our approach on a real hand-held device.

Among all the sensory abilities of human, vision is the most powerful one. As a matter of fact, vision can contribute to, influence or even replace the other sensory abilities. So, if intelligent machines are ever to be made, the computer vision domain has to play a very important role. To this end, faultless category-level recognition system is an essential need. Although, we are still far from this goal, the vision community is working hard to put the pieces of the puzzle together. We feel proud to be able to add our piece to that puzzle.

Chapitre 6

Conclusion et perspectives

6.1 Les contributions de ce travail

Au cours de cette thèse, nous avons travaillé sur la représentation d'images à partir d'information couleur et spatiale. Dans les chapitres 3 et 4, nous avons proposé de nouvelles représentations et analysé leur performance dans le domaine de la classification d'images en utilisant les sacs de mots. Par la suite, dans le chapitre 5, nous nous sommes penchés sur le problème de l'acquisition de données multi-spectrales à partir de systèmes d'acquisition bon marché comme les téléphones portables, les ordinateurs portables ou les tablettes numériques. Nous avons proposé une solution qui trouve de nombreuses applications dans le domaine de la vision par ordinateur.

Dans le chapitre 3, notre première contribution a été de montrer que les relations spatiales entre les mots visuels dans une image sont une information importante dans le contexte de la classification d'images. De plus, nous avons montré que cette information est complémentaire à celle fournie par les pyramides (SPR). Ce qui nous permet d'associer notre descripteur aux SPR et ainsi accroître la qualité des résultats de classification. Nous avons ainsi proposé un descripteur original appelé "Soft Pairwise Similarity angle distance histogram" (SPS_{ad}) qui pondère la contribution de chaque paire de motifs par la similarité entre des descripteurs de ces motifs. Nous avons montré que la distribution spatiale des motifs similaires dans une image est une information discriminante et peut améliorer significativement les résultats de classification. De plus, nous avons prouvé que la pondération de chaque paire par la similarité entre leurs descripteurs (soft similarity) est plus pertinente et plus robuste que de pondérer équitablement toutes les paires (hard similarity). En effet, notre descripteur SSP_{ad} augmente le taux de classification de 16% sur la base caltech101 et de 7% sur la base 15Scene par rapport à la représentation par sac de mots. Lorsque nous l'associons aux pyramides (SPR), notre descripteur offre de meilleurs résultats que les

SPR elles-mêmes ainsi que les autres descripteurs de la littérature fondés sur une combinaison avec les SPR. Dans ce cas, notre approche améliore le taux de classification de 3.5% sur Caltech101 et de 2.5% sur 15Scene par rapport aux SPR.

Dans le chapitre 4, nous avons proposé un nouveau descripteur couleur. Pour cela, nous avons décidé d'apprendre ce descripteur plutôt que de le définir à partir de modèles classiques, comme cela est fait dans la littérature. En effet, les approches classiques tentent de définir un descripteur couleur qui présente un certain degrés d'insensibilité à certaines variations radiométriques et photométriques. Dans notre cas, nous avons optimisé exclusivement le caractère discriminant de notre descripteur, l'invariance étant une conséquence de cette phase d'apprentissage. Dans cette thèse, nous avons présenté un descripteur spécifique et un descripteur générique. Le premier est appris sur les images d'apprentissage d'une base et testé sur les images de test de cette même base alors que le second est appris et testé sur des bases différentes. Nous avons constaté que les deux descripteurs permettent d'obtenir de meilleurs résultats que les descripteurs couleur de l'état de l'art. Nous avons aussi combiné notre descripteur avec les descripteurs SIFT et montré que nous obtenons de meilleurs résultats que l'état de l'art sur 2 bases d'images sur 4 et que nous sommes plus performants que les SIFT-couleur dans tous les cas.

Dans le chapitre 5, nous avons défini une nouvelle approche pour acquérir des données multi-spectrales à partir de systèmes d'acquisition bon marché. L'originalité de notre travail réside dans le fait que nous modifions la fonctionnalité des écrans d'affichage de ces systèmes pour qu'ils jouent le rôle de source de lumière. Dans ces conditions particulière, la source et le capteur étant des données intrinsèques au système d'acquisition, nous avons défini un algorithme qui exploite cette connaissance du système pour estimer les reflectances spectrales des objets observés. Nous avons ainsi montré que l'acquisition de données multi-spectrales peut être réalisée avec des systèmes d'acquisition grand-public et que l'acquisition exploitant les 3 sources de ces systèmes est plus performante que l'acquisition sous une unique source de lumière. Nous avons présenté des résultats sur des données synthétiques et des données réelles et dans les deux cas, nous avons amélioré les résultats des approches classiques. Nous avons ainsi montré que l'utilisation des 3 sources peut améliorer les résultats de 65% (RMSE) par rapport à l'acquisition sous une seul source et que notre algorithme de reconstruction permet de diminuer l'erreur de 16% par rapport à la méthode de Park [70].

6.2 Perspectives

Cette thèse offre de nombreuses perspectives. Dans le chapitre 3, les relations spatiales entre des motifs similaires ont permis d'améliorer significativement les résultats des sacs de mots. Il serait intéressant d'étendre notre approche à d'autres méthodes de codage comme le "sparse coding" [98] ou les "fisher vectors" [73] qui ont fait leur preuve ces dernières années. De même, la prise en compte de plusieurs informations comme la couleur et la forme pour les relations spatiales semble aussi une direction prometteuse.

Le descripteur couleur défini au chapitre 4 et destiné à être associé à un descripteur de forme, puisque cette association (forme-couleur) permet toujours d'améliorer les résultats de classification par rapport à la prise ne compte d'une seule information. Il apparaît donc opportun d'utiliser l'information du descripteur forme auquel sera associé notre descripteur couleur au cours de la phase d'apprentissage, permettant ainsi de mettre au point un descripteur le plus complémentaire possible. De plus, tout comme les descripteurs "Color Names" (CN) reposent sur des probabilités d'appartenance de chaque pixel à chacun des CN, nous envisageons d'exploiter la probabilité d'appartenance des pixels à chaque "cluster" couleur défini lors de notre apprentissage (soft assignment).

Notre travail sur l'acquisition multi-spectrale est très prometteur. Cependant, l'approche actuelle ne prend pas en compte la lumière qui éclaire la scène avant d'allumer l'écran du système d'acquisition. Comme l'obscurité totale n'est pas facile à obtenir, il serait intéressant de prendre en compte cette composante dans l'algorithme de reconstruction. Chi et al. [14] ont proposé une méthode pour ajouter cette information mais leur approche n'est pas applicable directement à notre cadre de travail à cause du faible nombre d'illuminants et de la faible intensité de l'écran de projection.

Parmi toutes les capacités sensorielles de l'être humain, la vision est la plus puissante. En effet, la vision peut contribuer, influencer ou même remplacer les autres capacités. Ainsi, si des machines intelligentes doivent être réalisées, le domaine de la vision par ordinateur jouera une large part dans cette conception. Dans ce contexte, les systèmes de reconnaissance ou de classification infaillibles font partie des besoins essentiels. Même si nous en sommes encore loin, la communauté travaille dure pour assembler les pièces du puzzle. Nous sommes fiers d'avoir contribuer à ce travail collectif.

Bibliography

[1] The barcelona calibrated images database, http://www.cvc.uab.es/color _ calibration/.

[2] Spectral database, university of joensuu color group, http://spectral.joensuu.fi/.

[3] A. Alsam and R. Lenz. Calibrating color cameras using metameric blacks. *Journal of the Optical Society of America*, 24:11–17, 2007.

[4] J. M. Alvarez, T. Gevers, and A. Lopez. Learning photometric invariance for object detection. *International Journal of Computer Vision*, 90(1):45–61, 2010.

[5] H. Bay, T. Tuytelaars, and L. J. V. Gool. Surf: Speeded-up robust features. In *European Conference on Computer Vision*, pages 404–417, 2006.

[6] R. Benavente, M. Vanrell, and R. Baldrich. Parametric fuzzy sets for automatic color naming. *Journal of the Optical Society of America*, 25(10):2582–2593, 2008.

[7] B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution.* University of California Press, Berkeley, CA, 1969.

[8] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *ACM International Conference on Image and Video Retrieval*, pages 401–408, 2007.

[9] R. Cavet, S. Volmer, E. Leopold, J. Kindermann, and G. Paaß. Revealing the connoted visual code: a new approach to video classification. *Computers & Graphics*, 28(3):361–369, 2004.

[10] Y. Chai, V. S. Lempitsky, and A. Zisserman. Bicos: A bi-level co-segmentation method for image classification. In *Computer Vision and Pattern Recognition*, pages 2579–2586, 2012.

[11] Y. Chai, E. Rahtu, V. S. Lempitsky, L. J. V. Gool, and A. Zisserman. Tricos: A tri-level class-discriminative co-segmentation method for image classification. In *European Conference of Computer Vision*, 2012.

[12] K. Chatfield, V. Lemtexpitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, pages 76.1–76.12, 2011.

[13] Q. Chen, Z. Song, Y. Hua, Z. Huang, and S. Yan. Hierarchical matching with side information for image classification. In *Computer Vision and Pattern Recognition*, pages 3426–3433, 2012.

[14] C. Chi, H. Yoo, and M. Ben-Ezra. Multi-spectral imaging by optimized wide band illumination. *International Journal of Computer Vision*, 86(2-3):140–151, 2010.

[15] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[16] E. A. Day, R. S. Berns, L. A. Taplin, and F. H. Imai. A psychophysical experiment evaluating the color accuracy of several multispectral image capture techniques. In *PICS*, pages 199–204, 2003.

[17] T. Deselaers and V. Ferrari. Global and efficient self-similarity for object classification and detection. In *Computer Vision and Pattern Recognition*, pages 1633–1640, 2010.

[18] I. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *JMLR*, 3:1265–1287, 2003.

[19] N. M. Elfiky, F. S. Khan, J. van de Weijer, and J. Gonzàlez. Discriminative compact pyramids for object and scene recognition. *Pattern Recognition*, 45(4):1627–1636, 2012.

[20] B. Fernando, E. Fromont, and T. Tuytelaars. Effective use of frequent itemset mining for image classification. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *European Conference on Computer Vision*, volume 7572 of *Lecture Notes in Computer Science*, pages 214–227. Springer, 2012.

[21] G. Finlayson and S. Hordley. Gamut constrained illumination estimation. *International Journal of Computer Vision*, 67(1):93–109, 2006.

[22] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[23] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *International Conference of Computer Vision*, October 2009.

[24] B. Funt and G. Finlayson. Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995.

[25] J. C. Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders. Kernel codebooks for scene categorization. In *European Conference of Computer Vision*, ECCV '08, pages 696–709, Berlin, Heidelberg, 2008. Springer-Verlag.

[26] J. Geusebroek, R. van den Boomgaard, A. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001.

[27] T. Gevers and A. Smeulders. Color based object recognition. *Pattern Recognition*, 32:453–464, 1999.

[28] T. Gevers and H. Stokman. Robust histogram construction from colour invariants for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):113–118, 2004.

[29] H. Haneishi, T. Hasegawa, A. Hosoi, Y. Yokoyama, N. Tsumura, and Y. Miyake. System design for accurately estimating the spectral reflectance of art paintings. *Applied Optics*, 39(35):6621–6632, Dec 2000.

[30] T. Harada, Y. Ushiku, Y. Yamashita, and Y. Kuniyoshi. Discriminative spatial pyramid. In *Computer Vision and Pattern Recognition*, pages 1617–1624, 2011.

[31] C. Harris and M. Stephens. A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151, 1988.

[32] V. Heikkinen, R. Lenz, T. Jetsu, J. Parkkinen, , M. H. Kasari, and T. Jääskelainen. Evaluation and unification of some methods for estimating reflectance spectra from rgb images. *Journal of the Optical Society of America - A*, 25(10):2444–2458, October 2008.

[33] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition*, CVPR '97, pages 762–768, Washington, DC, USA, 1997. IEEE Computer Society.

[34] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *International Conference on Computer Vision*, ICCV, pages 604–610, Washington, DC, USA, 2005. IEEE Computer Society.

[35] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. pages 506–513, 2004.

[36] F. Khan, R. Anwer, J. van de Weijer, A. Bagdanov, M. Vanrell, and A. Lopez. Color attributes for object detection. In *Computer Vision and Pattern Recognition*, 2012.

[37] F. Khan, J. Van de Weijer, A. Bagdanov, and M. Vanrell. Portmanteau vocabularies for multi-cue image representation. In *Neural Information Processing Systems*, 2011.

[38] F. S. Khan, J. van de Weijer, and M. Vanrell. Top-down color attention for object recognition. *IEEE Internet Computing*, pages 979–986, 2009.

[39] F. S. Khan, J. van de Weijer, and M. Vanrell. Modulating shape features by color attention for object recognition. *International Journal of Computer Vision*, 98(1):49–64, 2012.

[40] R. Khan, C. Barat, D. Muselet, and C. Ducottet. Spatial orientations of visual word pairs to improve bag-of-visual-words model. In *British Machine Vision Conference*. BMVA, 2012.

[41] R. Khan, J. van de Weijer, D. Karatzas, and D. Muselet. Towards multispectral data acquisition with hand-held devices. In *International Conference on Image Processing*, Melbourne, Australia, 2013.

[42] R. Khan, J. van de Weijer, F. S. Khan, D. Muselet, C. Ducottet, and C. Barat. Discriminative color descriptors. In *Computer Vision and Pattern Recognition*, Portland, USA, 2013.

[43] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, CVPR*, Colorado Springs, CO, June 2011.

[44] S. Kim, X. Jin, and J. Han. Disiclass: discriminative frequent pattern-based image classification. In *International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 7:1–7:10, New York, NY, USA, 2010. ACM.

[45] J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. 55:367–375, 1987.

[46] J. Krapac, J. J. Verbeek, and F. Jurie. Modeling spatial layout with fisher vectors for image categorization. In *International Conference of Computer Vision*, pages 1487–1494, 2011.

[47] J. Krapac, J. J. Verbeek, and F. Jurie. Spatial fisher vectors for image categorization. Technical report, INRIA Grenoble, 2011.

[48] P. Kubelka. New contribution to the optics of intensity light-scattering materials. *Journal of the Optical Society of America*, 38(5):448–457, 1948.

[49] J. H. Lambert. *Photometria*. Eberhard Klett, 1760.

[50] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

[51] B. Leibe, K. Mikolajczyk, B. Schiele, and T. Darmstadt. Efficient clustering and matching for object class recognition. In *British Machine Vision Conference*, 2006.

[52] F.-F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *Workshop on Generative-Model Based Vision*, 2004.

[53] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition*, volume 2, pages 524–531, 2005.

[54] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30:79–116, 1998.

[55] D. Liu, G. Hua, P. A. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *Computer Vision and Pattern Recognition*, 2008.

[56] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[57] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[58] M. R. Luo, G. Cui, and B. Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Colour Research and Application*, 26:340–350, 2001.

[59] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *Computer Vision and Pattern Recognition*, 2008.

[60] L. T. Maloney. Evaluation of linear models of surface spectral reflectance with small numbers of parameters. *Journal of the Optical Society of America*, 3(10):1673–1683, Oct 1986.

[61] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from. In *British Machine Vision Conference*, pages 384–393, 2002.

[62] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Computer Vision and Pattern Recognition*, pages 257–263, Madison, USA, 2003.

[63] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, Nov. 2005.

[64] F. Mindru, T. Tuytelaars, L. V. Gool, and T. Moons. Moment invariants for recognition under changing viewpoint and illumination. *Computer Vision and Image Understanding*, 94(1-3):3–27, Apr. 2004.

[65] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[66] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition*, CVPR '06, pages 2161–2168, Washington, DC, USA, 2006. IEEE Computer Society.

[67] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*, pages 490–503, 2006.

[68] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, May 2001.

[69] D. Parikh. Recognizing jumbled images: The role of local and global information in image classification. In *International Conference of Computer Vision*, pages 519–526, 2011.

[70] J. Park, M. Lee, M. D. Grossberg, and S. K. Nayar. Multispectral Imaging Using Multiplexed Illumination. In *International Conference of Computer Vision*, Oct 2007.

[71] J. P. S. Parkkinen, J. Hallikainen, and T. Jaaskelainen. Characteristic spectra of munsell colors. *Journal of the Optical Society of America*, 6(2):318–322, Feb 1989.

[72] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *European Conference of Computer Vision*, pages 464–475, 2006.

[73] F. Perronnin, J. Snchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference of Computer Vision*, 2010.

[74] J. Qin and N. H. Yung. Scene categorization via contextual visual words. *Pattern Recognition*, 43:1874–1888, 2010.

[75] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *Computer Vision and Pattern Recognition*, pages 2033–2040, 2006.

[76] S. Shafer. Using color to seperate reflection components. *Colour Research and Application*, 10(4):210–218, Winter 1985.

[77] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[78] R. Shreshtha, J. Y. Hardeberg, and R. Khan. On the design of multispectral color filter arrays. In *IS&T/SPIE Electronic Imaging*, 2011.

[79] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *Computer Vision and Pattern Recognition*, pages 1585–1592, 2011.

[80] Y. Su and F. Jurie. Visual word disambiguation by semantic contexts. In *International Conference of Computer Vision*, pages 311–318, 2011.

[81] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[82] P. Tirilly, V. Claveau, and P. Gros. Language modeling for bag-of-visual words image categorization. In *Conference on Image and Video Retrieval*, pages 249–258, 2008.

[83] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5), 2010.

[84] A. Torralba and A. Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE, 2011.

[85] T. Tuytelaars and C. Schmid. Vector quantizing feature space with a regular lattice. In *International Conference of Computer Vision*, pages 1–8, Rio de Janeiro, Brésil, 2007. IEEE Computer Society.

[86] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[87] J. van de Weijer, C. Schmid, J. J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18:1512–1523, 2009.

[88] E. Vazquez, R. Baldrich, J. van de Weijer, and M. Vanrell. Describing reflectances for color segmentation robust to shadows, highlights, and textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):917–930, 2011.

[89] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *International Conference of Computer Vision*, 2009.

[90] B. A. Wandell and J. E. Farrell. Water into wine: Converting scanner rgb to tristimulus xyz. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, pages 92–101. International Society for Optics and Photonics, 1993.

[91] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition*, 2010.

[92] J. V. D. Weijer and C. Schmid. Coloring local feature extraction. In *European Conference of Computer Vision*, pages 334–348, 2006.

[93] J. V. D. W. V. D. Weijer, T. Gevers, and A. Bagdanov. Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:150–156, 2005.

[94] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[95] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *International Conference on Computer Vision*, pages 1800–1807. IEEE Computer Society, 2005.

[96] L. Wu, M. Li, Z. Li, W. ying Ma, and N. Yu. Visual language modeling for image classification. In *Multimedia Information Retrieval*, pages 115–124, 2007.

[97] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *ACM Multimedia Information Retrieval Workshop*, pages 197–206, 2007.

[98] J. Yang, K. Yu, Y. Gong, and T. S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition*, pages 1794–1801, 2009.

[99] Y. Yang and S. Newsam. Spatial pyramid co-occurence for image classification. In *International Conference of Computer Vision*, 2011.

[100] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[101] J. Yuan, Y. Wu, and M. Yang. From frequent itemsets to semantically meaningful visual patterns. In *Knowledge Discovery and Data Mining*, pages 864–873, 2007.

[102] E. Zhang and M. Mayo. Improving bag-of-words model with spatial information. In *International Conference of Image and Vision Computing New Zealand*, 2010.

[103] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *Computer Vision and Pattern Recognition*, CVPR '06, pages 2126–2136, 2006.

[104] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object catergories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–218, 2007.

[105] Y. Zheng, H. Lu, C. Jin, and X. Xue. Incorporating spatial correlogram into bag-of-features model for scene categorization. In *Asian Conference on Computer Vision*, pages 333–342, 2009.

[106] G. Zhou, Z. Wang, J. Wang, and D. Feng. Spatial context for visual vocabulary construction. In *International Conference on Image Analysis and Signal Processing*, pages 176 –181, 2010.

[107] X. Zhou, K. Yu, T. Zhang, and T. Huang. Image classification using super-vector coding of local image descriptors. In *European Conference on Computer Vision*, 2010.